The Journal of Educational Psychology

Raymond G. Kuhlen, Editor
Syracuse University

Volume 49, 1958

Arthur C. Hoffman, Managing Editor
Helen Orr, Promotion Manager
Sarah Womack, Editorial Assistant

PUBLISHED BIMONTHLY BY
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1333 SIXTEENTH STREET N.W.
WASHINGTON 6, D. C.

CONTENTS OF VOLUME 49

THE JOURNAL OF

EDUCATIONAL PSYCHOLOGY

Volume 49 February, 1958 Number 1

EDITORIAL NOTE

This is the first issue of the first complete volume of the *Journal of Educational Psychology* to be published by the American Psychological Association and the first issue under a new editor. Delays and other matters incident to the transition of the journal to *both* a new publisher and a new editor have resulted in certain temporary departures from usual policy and some delay in the publication of first issues. Contributors and readers will note in this and the next several issues, for example, certain inconsistencies in style and certain departures from standard practice of publishing manuscripts in order of their receipt. In the future, contributors should follow the style prescribed by the *Publication Manual of the American Psychological Association.*

R. G. K.

Journal of Educational Psychology
Vol. 49, No. 1, 1958

PATTERNS OF PERSONAL PROBLEMS OF ADOLESCENT GIRLS[1]

RICHARD E. SCHUTZ[2]

Teachers College, Columbia University

Numerous investigators have determined the frequency with which various problems are listed by given samples of adolescents (4). However, in tabulating the problems it has been necessary to rely on a priori systems of classification. The customary procedure has been to categorize the problems by activity or functional areas; e.g., school, home, health, etc.

The purpose of the present study was to determine the pattern or structure underlying the personal problems which adolescents recognize and are willing to report on a youth problems inventory. This pattern was investigated by extracting homogeneous clusters from a sample of 156 items selected from the inventory.

Procedure

The inventory used in the study was the Billett-Starr Youth Problems Inventory, Senior Level (2), a check list intended to provide a means of systematically identifying the personal problems of individual adolescents. The 441 items which make up the Inventory include problems mentioned in the compositions and free responses of several large samples of high school students which the authors obtained in developing the instrument. The items are organized into 11 areas designated as follows:

1. Physical Health, Fitness, and Safety
2. Getting Along with Others
3. Boy-Girl Relationships
4. Home and Family Life
5. Personal Finance
6. Interests and Activities
7. School Life
8. Heredity
9. Planning for the Future
10. Mental-Emotional Health and Fitness
11. Morality and Religion

The cluster analysis was based on the responses of 500 girls in Grades 10 and 11 in two Pinellas County, Florida, high schools who took the Inventory as part of the national standardization program in May 1956. The schools are three-year high schools and had a segregated white enrollment at the time of the study. The Inventory was administered in regular classrooms by regular teachers. The Inventories were signed by the students.

The basic technique of analysis was that described by Loevinger, Gleser, and Dubois (6) for deriving clusters which have maximum reliability as estimated by Kuder-Richardson Formula 20. Each cluster is obtained by starting with a triad of items having the highest covariance and adding items in succession, adding always the item for which the ratio of the sum of covariance with the items already in the cluster is a maximum. Items are added to the cluster until no more items remain which will increase this ratio. The process is repeated on the residual pool of items to form the second and subsequent clusters.

[1] Based on a dissertation submitted in partial fulfillment of the requirements for the Ph.D. at Columbia University. The investigation was carried out under the helpful direction of Robert L. Thorndike. The writer wishes to thank Roger T. Lennon and the staff of the Division of Test Research and Service, World Book Co. for making the study possible.

[2] Now at Arizona State College, Tempe, Arizona.

A sample of 156 items was selected to be included in the cluster analysis (**8**, pp. 57–64). The items selected have the following characteristics: (*a*) they are included in both the Junior and Senior levels of the Inventory; (*b*) they were rated as "very serious" or "moderately serious" problems by a panel of 20 guidance specialists; (*c*) each was marked by at least 5% of the *S*s in the sample. An attempt was made to make the sample of items representative of the 11 areas of the Inventory and to include as many items as possible which have counterparts in other published problems check lists.

The Inventory attempts to get at the intensity of a student's problems by allowing him to differentiate between those which bother him "some" and those which bother him "very much." For the present analysis each *S*'s "some" and "much" responses were combined into a single category.

Each *S*'s responses were multiple punched on an IBM card, and the 156 by 156 co-occurrence matrix was prepared using the counting sorter. The figures in the co-occurrence matrix were converted to percentages and the variance-covariance matrix prepared. The cluster analysis was performed, and a check was made on the factorial purity and reliability of the obtained clusters.

Results

Three clusters were extracted from the pool of 156 items. Eighty-three items are included in Cluster I, 16 in Cluster II, and 17 in Cluster III. An abbreviated Cluster I, consisting of 37 items, was formed by eliminating 35 items which correlated less than .40 with the complete cluster and 11 items which nearly duplicated another item in the cluster; e.g., the two items, "I'm often restless," and "I'm restless most of the time." The items in the abbreviated Cluster I and in Clusters II and III are shown in Tables 1–3. The items are arranged in order of the magnitude of the point biserial correlation of each item with its cluster. The area and item number within the area are indicated in the first column of each table. The Kuder-Richardson Formula 20 reliabilities of the clusters and their intercorrelations are shown in Table 4.

The nature of a cluster must be determined by examining the items to discover the general attribute they seem to hold in common. The items in Cluster I cover a broad area, coming from eight areas of the Inventory. The cluster appears to reflect a general feeling of personal anxiety and insecurity.

The items in Cluster II are currently classified under seven different area headings in the Inventory. They seem to involve a feeling of nervous tension concerning relationships with other persons. This cluster is the least homogeneous of the three, and its correlation with Cluster I is nearly as high as its reliability. Both Clusters I and II reflect personal anxiety. While the items in Cluster II did not have enough in common with the items in Cluster I to be included in the more homogeneous general cluster, they shared sufficient common variance to form another with lower reliability.

Cluster III is the only cluster that does not cut across the functional area organization of the Inventory to any great extent. Fifteen of the 17 items in the cluster come from Area IV of the Inventory headed "Home and Family Life." The items all represent some kind of difficulty in getting along with parents.

If a cluster is factorially pure, all of its common factor variance should be accounted for by a single centroid factor. The ratio of the first factor variance to the common factor variance thus provides a basis for evaluating the factorial purity of a cluster.

A complete centroid analysis (**9**) was performed independently on the items in the abbreviated Cluster I and in Clusters II and III. The highest correlation coeffi-

TABLE 1

CLUSTER I

Item No.	r	Item
10–38	.59	People don't understand me.
10–10	.57	I'm afraid of making mistakes.
10–15	.55	I'm often restless.
10–43	.51	I worry about what others say.
10–37	.51	I feel I'm not wanted.
10–17	.50	I'm disgusted with myself (dislike myself very much).
10–19	.49	I need someone to give me advice.
8–3	.49	I would like to be able to do something well.
8–1	.49	I don't understand myself.
11–10	.49	Many times I don't know what is right and what is wrong.
10–39	.48	People talk about me behind my back.
10–2	.48	I feel uncertain (unsure) about everything.
10–34	.47	I spend too much time daydreaming.
10–3	.46	I need to learn to depend on myself.
10–35	.45	I feel sorry for myself.
10–46	.45	I get excited too easily.
11–6	.45	I'm sometimes troubled by immoral (bad) thoughts.
7–5	.45	I wonder if I'll pass.
10–32	.44	I don't get out and go after what I want.
9–18	.44	I wonder if I'm taking the right subjects.
7–44	.44	I'm afraid to take tests.
10–58	.44	I'd like to know how to get rid of a bad habit.
4–46	.44	I'm unhappy at home.
7–62	.43	Some teachers never encourage or help me.
6–1	.43	I seldom have anything interesting to do.
7–85	.43	I would like to know how to get along with certain teachers.
8–2	.43	I wonder what my real mental ability is.
10–13	.42	I wonder what my future will be.
10–1	.42	I'm confused by the way things change.
2–50	.42	I feel lonely most of the time.
4–40	.42	I'm afraid to tell my (father) (mother) when I've done something wrong.
10–42	.42	I'm blamed for things that aren't my fault.
2–52	.41	I find it hard to make friends.
10–30	.41	I don't know how to (pay attention) (work or study hard).
11–5	.41	I often tell lies.
10–45	.40	I'm bothered by people who find fault with me.
10–47	.40	I can't control my temper.

cient in each column of the matrix was used as the communality estimate, communalities being re-estimated by this method for every residual matrix. Factoring was considered complete when both Tucker's Phi and Coombs' criterion (**3**) indicated that the last significant factor had been extracted. Three factors were extracted from Cluster I, three from Cluster II, and two from Cluster III. First factors account for 80%, 69%, and 85%, respectively, of the common factor variance of the clusters.

To investigate the extent to which the reliability of the clusters is dependent on chance factors, the Kuder-Richardson Formula 20 reliability of each cluster was computed, based on the results of a new sample of 73 *S*s selected from the same population as the original sample. The reliability coefficients for the new sample were as follows: Cluster I, .94; abbrevi-

TABLE 2

CLUSTER II

Item No.	*r*	Item
2–11	.63	I'm nervous when I talk to people.
2–15	.60	I'm not good at talking with people.
7–38	.55	I'm nervous in front of the class.
1–19	.49	I get tired easily.
2–10	.48	(I'm afraid) (I don't like) to meet people.
2–39	.47	I want others to like me.
1–21	.44	I'm always nervous.
3–1	.44	I don't understand (boys) (girls).
2–6	.38	I'm not good-looking.
6–10	.36	I would rather be alone.
6–9	.36	I get tired from too much activity.
6–11	.36	I spend too much time on (radio) (television) (movies).
1–22	.32	I need to know more about sex (body changes at my age) (new body functions).
10–40	.31	I'm afraid I seem conceited (stuck-up).
9–25	.30	I'm not sure whether I should go to college.
2–7	.28	I (don't have) (don't know how to pick) the right clothes.

TABLE 3

CLUSTER III

Item No.	*r*	Item
4–31	.64	My (father) (mother) is always criticizing (blaming) (nagging) me.
4–43	.59	I can't discuss things with my (father) (mother).
4–33	.58	My (father) (mother) is always expecting too much of me.
4–37	.54	I'm often the cause of family quarrels (parents argue about things I do).
4–51	.52	I'm thinking of leaving home.
4–41	.52	My (father) (mother) has little or no interest in what I do.
4–39	.51	I sometimes lie to my (father) (mother) to get permission to do something.
4–60	.51	My (father) (mother) is often nervous and irritable.
4–42	.50	My (father) (mother) never asks my opinion about anything important to the family.
4–50	.49	I dislike my (father) (mother) very much.
4–25	.49	I don't agree with my (father) (mother) about my out-of-school activities.
4–48	.48	I'm sometimes ashamed of things my parents do or say.
4–38	.45	My (father) (mother) pries into my private affairs.
4–26	.44	My (brother) (sister) is always causing me trouble.
5–4	.43	I don't get an allowance.
2–19	.40	I often "stretch the truth" when I tell something.
4–49	.34	There's too much drinking in our home.

ated Cluster I, .91; Cluster II, .63; Cluster III, .83.

The significance of the difference between each of the coefficients for the new and the original group was computed using Fisher's *z* transformation. The difference for Cluster II is significant at the .001 level. There is no significant difference

TABLE 4
Cluster Intercorrelations and Kuder-Richardson Reliabilities

Cluster	I(83)	I(37)	II	III
I (83)	(.94)	.98	.67	.60
I (37)	.98	(.90)	.63	.52
II	.67	.63	(.71)	.34
III	.60	.52	.34	(.81)

between the coefficients for the other clusters.

Discussion of Results

The findings of the study have implications for both the theory and measurement of adolescent problems. While the study was not designed to test the validity of the various theories of adolescent problems that have been proposed **(1, 5, 7)**, it does provide evidence concerning the way in which problems cluster or "go together."

The cluster structure does not correspond closely to any of the theoretical frameworks which have been proposed for classifying adolescent problems. The composition of Cluster I suggests that a single dimension of personal anxiety underlies many of the manifest problems which theorists have used several dimensions to explain.

The fact that two of the clusters cut across several areas of the Inventory indicates that the classifying of items in problems check lists into the traditional functional or activity categories is in large part an arbitrary procedure. However, even though the conventional rubrics do not all represent true functional unities, teachers and counselors may still find the categories helpful. The subdivisions often suggest programs of action related to the kinds of services that schools and other social agencies are equipped to provide. Thus, the traditional area organization often serves the purpose of suggesting foci of therapeutic action.

Although no data are yet available concerning the predictive validity of the clusters, they provide a potential means of screening adolescents in need of psychological help. They have high face validity and are reasonably pure factorially. A study of predictive validity is required to determine the extent to which they are actually related to personal adjustment.

Summary

A cluster analysis was performed on 156 selected items from the Billett-Starr Youth Problems Inventory, Senior Level, based on the responses of 500 adolescent girls. Three clusters were extracted and designated as follows: Cluster I—General personal anxiety and insecurity; Cluster II—Tension concerning relations with others; and Cluster III—Difficulties in getting along with parents. The implications of the cluster structure for the theory and measurement of adolescent problems was briefly discussed.

REFERENCES

1. Ausubel, D. P. Problems of adolescent adjustment. *Nat. Assoc. sec. Sch. Prin. Bull.*, 1950, **34** (167), 1–84.
2. Billett, R. O. & Starr, I. S. *Billett-Starr Youth Problems Inventory.* Yonkers-on-Hudson, N. Y.: World Book, 1956.
3. Fruchter, B. *Introduction to factor analysis.* New York: Van Nostrand, 1954.
4. Grant, B. Survey of studies on problems of adolescents. *Calif. J. sec. Educ.*, 1953, **28,** 293–297.
5. Havighurst, R. J. *Human development and education.* New York: Longmans, Green, 1953.
6. Loevinger, Jane, Gleser, Goldine C. & Dubois, P. H. Maximizing the discriminating power of a multiple-score test. *Psychometrika*, 1953, **18,** 309–317.
7. Luchins, A. S. On the theories and problems of adolescence. *J. genet. Psychol.*, 1954, **85,** 47–63.
8. Schutz, R. E. Patterns of personal problems of adolescent girls. Unpublished doctoral dissertation, Columbia Univer., 1957.
9. Thurstone, L. L. *Multiple factor analysis.* Chicago: Univer. of Chicago Press, 1947.

Received August 10, 1957.

Journal of Educational Psychology
Vol. 49, No. 1, 1958

READABILITY LEVEL AND DIFFERENTIAL TEST PERFORMANCE: A LANGUAGE REVISION OF THE STUDY OF VALUES

JEROME LEVY[1]
Denver University

Psychologists have recently become interested in the "readability" of material; that is, how the verbal difficulty of written material affects the communication of ideas. Hebb and Bindra (**5**), Ogdon (**7**), and Stevens and Stone (**8**) have published material relevant to this question. These writers have been concerned with readability in discursive writings; e.g., textbooks. The important question, however, of how the readability level of a psychological test might affect the score which a subject with a defined level of verbal competence achieves has not hitherto been directly investigated. It appears to the writer that this is a question of considerable importance to sound psychometric practice. The present study was designed to investigate this question.

The Allport-Vernon Study of Values, 1951 Revision (**1**), was selected as an appropriate instrument. Since its introduction in 1931 and following its revision in 1951, the Study of Values has been employed in researches in personality theory, and for clinical evaluation and guidance purposes, and has proved quite useful in these areas. Many writers have noted, however, that the level of language employed in the Study of Values is quite complex and difficult. Gough (**4**), for example, writes: "The language used is too academic and involved for use in groups very far removed from a scholastic environment." Allport notes: "The scale is designed primarily for use with college students or with adults who have had college (or equivalent) education" (**1**). Although it is not explicitly stated, one surmises that the basis of this view is the recognition of the verbal complexity of the test, particularly in terms of vocabulary usage. If, however, verbal difficulty is the limiting factor as seems suggested, then the Study of Values may not be a valid test even for *all* college students. Classes in remedial reading, basic communication, and other related areas in every university attest to the large numbers of students deficient in these verbal skills.

The present study represents an attempt to produce a modification of the Study of Values congruent in all respects with the 1951 Revision, except that it employs less difficult language. Such a modification may be useful in two ways: if it can be demonstrated equivalent to the 1951 Revision it may be validly used with *S*s for whom the 1951 Revision is valid, and may also permit extension of the test to populations hitherto considered inappropriate (i.e., noncollege level *S*s). Secondly, it may be utilized in demonstrating the effects of differential vocabulary ability on test scores.

Problem

There are four major phases to this research: (*a*) actual construction of the modification; (*b*) a test of the meaning equivalence of the items of the modification and the 1951 Revision; (*c*) a demonstration that the language level of the modification is indeed lower than that of the 1951 Revision; and (*e*) a demonstration that there will be differences in performance between the two forms for *S*s

[1] Now Chief Psychologist, Division of Mental Health, Colorado State Department of Public Health.

whose vocabulary ability is below that inherent in the 1951 Revision which will be *significantly greater* than the between-form differences for *S*s whose vocabulary ability is equal to that of the 1951 Revision. This is the major focus of the investigation and the primary hypothesis to be tested. Consequent hypotheses for testing are: (*a*) For a high vocabulary ability group the between-form performance differences will be of a minimal nonsignificant nature; (*b*) For a low vocabulary ability group there will be a large and statistically significant difference in performance on the two forms; and (*c*) For a group of *S*s intermediate on the vocabulary criterion, the between-form differences will fall between that for the high and that for the low groups.

METHOD

Construction of the modification. Of the 45 items in the 1951 Revision, all but one were reworded. Items were constructed which attempted to maintain the structure and meaning of the equivalent original items, but which were worded much more simply. Where a choice of words with equivalent meanings was to be made, the simpler word was always chosen.

Equivalence of meaning. Both the 1951 Revision and the modified form were given to a group of judges with a rather heterogeneous background in psychology. These judges were unfamiliar with the Study of Values. They were supplied with the definitions of the six predominant personality types as they appear in the Manual of Directions (**1**) so that they would have clear referents for each scale. They were presented the two forms in mixed order and asked to judge which type of person would answer each question with each alternative choice. For example, in the item "Assuming you have sufficient ability, would you prefer to be: (a) a banker; (b) a politician?" which of Allport's types would choose alternative (a), which would choose (b)? The judges also matched the items of both forms by number; the items of each form having been randomized.

Readability. In order to evaluate the hypothesis that the language level of the modified form was simpler, a test of "readability" was employed. This is the Flesch formula (**3**).

Between-form differences. A group of 157 young male Air Force *S*s was administered the 1951 Revision, the present modification, and the Diagnostic Reading Test, Survey Section (**2**). The two forms of the Study of Values were given in balanced order of administration, with the Diagnostic Reading Test coming between them for all *S*s. The vocabulary scale of the Survey Section was taken as the primary criteria for the selection of three groups: a *Low* group of 22 *S*s whose vocabulary scores placed them at or below the lower quartile of the eighth grade, a *High* group of 23 *S*s with vocabulary scores at or above the upper quartile of the twelfth grade, and a *Middle* group of 20 *S*s with intermediate vocabulary scores. Thus there resulted three groups of approximately equal size, with no overlap on the vocabulary criterion, whose tested vocabulary ability closely corresponded to the readability of the two forms of the Study of Values. The two forms were scored for each *S*, and a "deviation score" was also computed. This is merely the difference between *S*'s score on a scale on one form and his score on the same scale on the alternate form, either positive or negative (sign being dropped), and summed for all six scales.

RESULTS

The results of the judgments indicate that, in the opinion of these judges, the revised items *do* ask the same thing as the original items. The judges made the same categorizations of the original and the revised items; i.e., there was not a statistically significant difference between

TABLE 1
SUMMARY OF t TESTS OF ORDER FOR THE HIGH, MIDDLE AND LOW GROUPS

Theoretical	Economic	Aesthetic	Social	Political	Religious
Low Group					
t[a] .11	1.12	.65	.08	.29	.07
Middle Group					
t[b] .98	1.07	.11	.44	.80	1.42
High Group					
t[c] .64	.26	.52	.58	.74	.59

[a] .05 = 2.086, df = 20
[b] .05 = 2.101, df = 18
[c] .05 = 2.080, df = 21

the type judgments on the pairs of items. The results of the item-to-item match show that, of the total of 506 matches the judges made, 495—or 98%—were matched correctly.

The Flesch formula yielded a "Reading Ease Score" of 52.7 for the 1951 Revision, placing it at the twelfth grade level: the score of 72.6 for the modified form is at the seventh grade level.

TABLE 2
ANALYSIS OF VARIANCE TO DETERMINE SIGNIFICANCE OF BETWEEN-FORM DIFFERENCES FOR THE LOW GROUP, AS COMPARED TO BETWEEN-FORM DIFFERENCES FOR THE HIGH GROUP

	Sum of Squares	df	Mean Square	F
Between	609.158	2	304.579	3.955
Within	4,774.780	62	77.103	
Total	5,383.938	64		

	Low Group	Middle Group	High Group
Mean Differences	24.182	22.150	17.043

Since some of the Ss in each of the three groups took the forms in original-modified order, and others took them in modified-original order, t tests were computed to determine whether the order in which the forms were taken has a significant effect on the between-form differences. As may be seen from Table 1, none of these t's was significant at the .05 criterion level. It was therefore possible to combine all of the Ss within each of the three groups, and to treat each of the groups as a single unit for further statistical evaluation.

As a test of the main hypothesis—that is, that the Low group would show significantly greater between-form differences than would the High group—an analysis of variance was performed utilizing the deviation scores. The F value of 3.955 which was obtained, as shown in Table 2, is significant at, and beyond, the .05 level of confidence. This clearly substantiates the major hypothesis. The primary characteristic on which these two groups differ is that of diagnosed vocabulary ability, and the significantly greater between-form differences in the

Low group seems clearly related to their low vocabulary level. The group means of the difference scores, which are also contained in Table 2, bear out the hypothesis that as the more difficult form moves away from the *S*s' level of vocabulary ability they make increasingly different scores than they do on the form which is within their level of competence.

The groups show a good deal of consistency in mean scale values from one form to another, as indicated by the data in Table 3. On the Theoretical scale, the Low group scores lowest on both forms, the Middle group is intermediate on both forms, and the High group scores highest on both forms. The Aesthetic scale shows the reverse order, with the same relative consistency on both forms. The *t* tests were computed to determine the significance of change in performance on each of the six scales for the High and Low groups: these data are presented in Table 4.

The hypotheses employed in this investigation state, in effect, that the two forms of the Study of Values are parallel

TABLE 3
MEAN SCALE VALUES ACHIEVED BY THE LOW, MIDDLE AND HIGH GROUPS ON THE 1951 REVISION AND THE MODIFICATION

Group	Theoretical	Economic	Aesthetic	Social	Political	Religious
			Low			
1951 Revision						
Mean	43.136	40.272	34.181	40.454	42.772	39.181
Modified Mean	40.136	42.363	33.954	41.045	42.454	40.045
			Middle			
1951 Revision						
Mean	44.500	43.550	32.350	35.250	43.350	41.000
Modified Mean	44.950	42.150	33.600	36.850	41.300	41.150
			High			
1951 Revision						
Mean	49.347	44.826	30.913	33.173	40.260	41.378
Modified Mean	47.739	44.304	31.478	35.478	40.260	40.739

TABLE 4
t TESTS OF SIGNIFICANCE OF DIFFERENCES IN PERFORMANCE BETWEEN 1951 REVISION AND MODIFICATION ON EACH OF THE SIX SCALES FOR HIGH AND LOW GROUPS

	Theoretical	Economic	Aesthetic	Social	Political	Religious
			Low Group			
t	3.86	2.29	.37	1.36	1.13	1.77
P	.001	.05	NS	NS	NS	NS
			High Group			
t	1.47	1.18	1.55	2.62	.46	1.97
P	NS	NS	NS	.02	NS	NS

TABLE 5

Test-Retest Correlation Coefficients for the 1951 Revision and Obtained Correlation Coefficients Between Performance on the 1951 Revision and the Modified Form for the High and Low Groups

	1951 Revision (Test-Retest)	Low Group correlations between 1951 Revision and Modified	High Group correlations between 1951 Revision and Modified
Theoretical	.87	.60	.76
Economic	.92	.53	.84
Aesthetic	.90	.60	.90
Social	.77	.74	.68
Political	.90	.55	.89
Religious	.91	.72	.91

Tests of Significance Between Correlation Coefficients

	High to Low Critical Ratio	High to 1951 Revision Critical Ratio	Low to 1951 Revision Critical Ratio
Theoretical	1.27	1.35	2.68[a]
Economic	2.65[a]	1.47	4.18[c]
Aesthetic	3.27[b]	0.00	3.26[b]
Social	.51	.76	.79
Political	3.18[a]	.30	3.57[b]
Religious	2.60[a]	0.00	2.59[a]

[a] .01 level
[b] .001 level
[c] .0001 level

forms for the High vocabulary group where there are not the interfering effects of vocabulary limitations, and are not parallel forms for the Low group where these limitations do exist. If the two forms are parallel, the correlation between performance on the Modification and on the 1951 Revision should not differ significantly from a test-retest correlation of performance on the 1951 Revision alone. In other words, if the two forms are equivalent there should not be a statistically significant difference between the correlation coefficients obtained by giving the 1951 Revision twice, and those obtained by correlating performance on the two forms. Pearson product-moment correlation coefficients were computed to determine the degree of relationship of the two forms in the High and Low groups. To determine whether these *r*'s are significantly different from each other the r-to-z transformation and critical ratio test of significance was employed. These data are presented in Table 5. Because of the impossibility of obtaining test-retest data on this group of *S*s, these data are taken from the 1951 Revision Manual (**1**).

Discussion

It seems reasonable to expect that people who are highly theoretically oriented are unlikely to have a very low vocabulary level. Omitting any question of cause and effect, it does seem likely that there is a positive relationship between these areas. People who are theoretically oriented are likely to spend a good deal of time reading in many areas of knowledge, and in other ways are apt to develop a well-organized and comprehensive vocabulary. Conversely, persons not theoretically oriented may reasonably be expected to engage less in these activities. What the writer proposes, in effect, is that the higher mean Theoretical score found on the original form is *spuriously high*. When items are not understood, choices are likely to be made on the basis of something other than true preference, and it is suggested that the higher mean for this group on the original form is due largely to noncomprehension. When, on the other hand, they are presented the same alternative choices in language that they can understand, they tend to score lower. Since as a group they are not highly theoretically oriented, the group mean score on this scale will fall as the meaning of the items become clear. If this view is correct, then the modified form is indeed a more valid measure of theoretical values in this group.

The lower Economic scale scores for the Low group on the 1951 Revision is significant at the .05 level. It is suggested that this result again reflects the effects of noncomprehension. It is hypothesized that when choices having to do with economic values are presented to *S*s in the Low group in language that is meaningful to them, they score higher on the Economic scale because this reflects a "true" underlying orientation in this direction.

The results shown in Table 4 also indicate that there is one scale on which the High group *S*s make significantly different scores on the two forms: the Social scale. Two tentative explanations may be advanced for the significantly higher score on the modified than on the original form: (*a*) this may be a chance phenomenon, or (*b*) real differences may arise on this scale in a highly verbal group as a result of change in wording. A cross-validational study to investigate the meaning of this result is necessary.

Interpretation of the results of this study seems simple and straightforward. Since none of the between form correlations for the High group differs significantly from the test-retest correlation coefficients, the modified form is considered a parallel form for this group. For those *S*s where there are not the vocabulary limitations, the modification yields results that do not differ significantly from that which would be obtained through the use of the 1951 Revision, and the two forms may be considered equivalent for this group. However, the significant CRs obtained in the comparison for the Low group indicate that the modification is *not* a parallel form for this group. Since the major characteristic on which the groups differ, and the basis on which they were selected, is differential vocabulary ability, it seems clearly evident that it is the vocabulary proficiency of the *S*s which determines differential performance on the two forms. If the *S*s can understand the language, then the modification is a parallel form of the Study of Values. If, however, the *S*s' language ability is not sufficient to cope with the 1951 Revision, the differences in performance on the two forms are so great as to indicate that they represent different tasks.

Since the modification is a parallel form for highly verbal *S*s, it is likely that it is as valid for these *S*s as is the 1951 Revision. It is the writer's opinion that the modification is a *more valid* form for low verbal people. Since the level of language is brought within the comprehension of these low verbal *S*s, random response choices based on noncomprehension are likely to be decreased. Responses based on an understanding of what the question is really asking are more likely to reflect truly the *S*'s orientation on the dimension being measured. External validation studies with different occupational groups, with other test criteria, and with generalized measures of personality are necessary for an empirical demonstration of validity. If on further investigation this modification receives external validity support, it may well serve a useful function in psychological evaluation. It will provide the same sort of personality information as does the 1951 Revision for highly verbal Ss and will permit extension of the test to populations hitherto considered inappropriate.

The factor of verbal level as an important variable in pencil-and-paper psychological test results has, of course, been realized for some time. But this has been at a rather gross level. This is the first study, to this writer's knowledge, which has attempted to manipulate verbal facility as an experimental variable and to evaluate statistically the effect which this has on test performance. With the growing concern about readability and communication in general, it is not likely to be the last. A good deal more study remains to be done with the modification. An item analysis is planned for the future. A covariance analysis of the relationship between differential test performance, verbal level, and intelligence is likely to clarify other important variables. But a promising start has been made in producing a meaningfully equivalent, yet less verbally complex, form of this popular test, and it has been possible to demonstrate the significant role which verbal facility plays in pencil-and-paper test results.

Summary

This paper reports an attempt to produce a meaningfully equivalent, but less verbally complex, form of the Study of Values, and to demonstrate the significant role which language facility plays in determining score patterns on pencil-and-paper test of personality. Pilot data with the modified form indicates that it is judged as asking the same things as the Study of Values. Flesch counts of the 1951 Revision and of this modification indicate that they are at the twelfth grade level and seventh grade level, respectively. Three experimental groups were established: a Low vocabulary group, a High vocabulary group, and a Middle group. The two forms were administered to these Ss. Analysis of the data indicates that for the High group the modification is an equivalent form. The Ss in the Low group make significantly different scores on the two forms, and the forms are not equivalent for this group. The differences in performance in this group are attributed to a vocabulary level inadequate to deal with the 1951 Revision, and it is suggested that for these Ss the modification provides a more valid test of value orientations. Much additional exploratory work remains to be done with the modification before it may be completely accepted as a valid, equivalent form for all Ss. This study serves to emphasize the important role of language facility in all pencil-and-paper personality tests.

REFERENCES

1. Allport, G. W., Vernon, P. E., & Lindsey, G. *Study of Values (Rev. ed.): Manual of Directions*. Boston: Houghton Mifflin, 1951.
2. Committee on Diagnostic Reading Tests, Inc. Description of the purposes and functions of the Diagnostic Reading Tests. *Educ. psychol. Measmt.* 1948, **8,** 3–14.
3. Flesch, R. *How to test readability*. New York: Harper, 1951.
4. Gough, H. G. In O. K. Buros (Ed.), *The fourth mental measurements yearbook*. Highland Park, New Jersey: Gryphon Press, 1953. Pp. 156–157.
5. Hebb, D. C., & Bindra, D. Scientific writing and the general problem of communication. *Amer. Psychologist*, 1952, **7,** 569–573.
6. Levy, J. Reducing the language complexity of the Study of Values: A revision. Unpublished doctoral dissertation, Univer. of Denver, 1956.
7. Ogdon, D. P. Flesch counts of eight current tests for introductory psychology. *Amer. Psychologist*, 1954, **9,** 143–144.
8. Stevens, S. S. & Stone, C. Psychological writing, easy and hard. *Amer. Psychologist*, 1947, **2,** 230–235: Comment 523–525.

Received July 12, 1957.

Journal of Educational Psychology
Vol. 49, No. 1, 1958

THE ATTITUDES STUDENTS ASSIGN TO THEIR TEACHER

NORMAN M. CHANSKY
Oswego State Teachers College

How students see and judge their teachers are behavioral operations which have more than theoretical importance. The validity of student judgments has never been determined, yet research studies in this area tacitly assume the factor of validity. That different students viewing the same instructor assign him varied, even contradictory, attitudes warrants inquiry into the factors associated with such judging. The purpose of the present study was to examine whether attitudes assigned to a teacher can in any way differentiate between students with authoritarian outlooks and those with democratic outlooks.

Procedure

After lecturing for three weeks about prenatal development, postnatal development, and theories of child development, the writer administered the Minnesota Teacher Attitude Inventory (MTAI) **(3)** to his classes in Child Psychology. The MTAI is a measure of attitude toward democratic attitudes and practices in teaching, based on the *F* Scale of the authoritarian personality series of studies **(1)**. High scores on the MTAI indicate a democratic attitude; low scores, an authoritarian attitude.

After having answered the items of the MTAI, the students were asked to write on the back of their answer sheets what attitudes toward children they thought their instructor held. Care had been taken during the three introductory weeks of the course to avoid controversial issues in child psychology. At no time did the instructor consciously cue the students to his point of view.

During the subsequent 12 weeks of the course, lectures and course activities centered on (*a*) the law of effect as related to the development of behavior in children, (*b*) factors associated with acceptance of children, and (*c*) the childhood antecedents of self realization.

At the end of the course, the MTAI was readministered and the students were asked once again to state the attitude toward children their instructor holds.

Results

A content analysis of the first set of student statements of the attitude which the instructor was presumed to hold yielded seven distinct categories. These categories were:

(1) freedom of children to manipulate the environment

(2) development of socially precise behaviors in children with punishment for deviants (discipline)

(3) a cold, impersonal attitude toward children

(4) the development of independence in children

(5) respect for children

(6) warm, friendly, personal relationship with children

(7) helping children with learning and emotional problems (clinical). Some attitudes could not be classified and were not included in the results.

The specific statistical tests (See Table 1) indicate that those students who felt their teacher would encourage freedom in children received significantly higher MTAI scores than any other group. On the other hand, those students who saw their instructor taking a clinical attitude toward children received significantly lower MTAI scores than most of the other groups. Their scores were not reliably different from either the discipline or the impersonal groups. In addition, the group who

TABLE 1
t RATIOS FOR THE HYPOTHECATED INSTRUCTOR ATTITUDES BEFORE EXPOSURE TO TEACHER ATTITUDE

Instructor Attitude	Number	Mean MTAI score	*t* Ratios for differences between attitude groups					
			Discipline	Impersonal	Independence	Respect	Personal	Clinical
Freedom	16	53.25	4.83[b]	2.19[a]	2.39[a]	2.27[a]	2.21[a]	4.33[b]
Discipline	12	15.42		−1.81	−5.02[c]	1.86	−3.45[b]	0.78
Impersonal	9	31.77			0.77	0.02	−0.70	1.70
Independence	12	38.04				1.04	0.12	3.50[c]
Respect	12	31.50					−0.75	3.59[c]
Personal	6	37.50						3.15[b]
Clinical	9	14.66						

Note.—degrees of freedom are $N_1 + N_2 - 2$.
[a] Significant at *P* .05 level.
[b] Significant at *P* .01 level.
[c] Significant at *P* .001 level.

looked upon their instructor as a disciplinarian received significantly lower MTAI scores than the freedom, independence, and personal groups.

These statistical tests indicate that, in the absence of well-defined cues, students projected their own attitudes toward the instructor (**2**). The reasoning behind this interpretation was as follows. Since cues to the attitude of the teacher toward children were minimized, the necessary ambiguity for projection was present. Furthermore, when those students who assigned such attitudes as freedom, independence, respect, usually associated with democracy, received higher MTAI scores, indicative of a democratic attitude, and when those students who perceived their instructor as either a disciplinarian or a clinician received lower MTAI scores, indicative of an authoritarian attitude, the interpretation of projection was confirmed. The relationship between discipline and authoritarianism is not unexpected, but why a service attitude such as that which a clinician holds received such low MTAI scores can be reconciled only if the clinical attitude represents an expression of the deep dependency needs of the authoritarian personality. If this hypothesis be substantiated in subsequent studies, it may be worthwhile investigating whether this *clinical* attitude means sensitivity to the needs of children or if it is a subtle way of expressing the wish for an anaclitic relationship with an authority figure.

Data obtained *after* exposure to the teacher's attitude are reported in Table II. Four new categories appear in the content analysis of the student assigned attitudes of the teacher. These are patience, school learning, no punishment, and understanding. A chi-square test was made to determine whether students assigned different attitudes to their instructor after he had presented his point of view. The data yielded the non-significant chi-square or 10.91 (10 *df*). Any change that took place can be attributed to chance. A coefficient of contingency of .46 was also obtained. These data further suggest that there was some consistency in responding to the teacher-attitude question on the two testings even after exposure to his attitude.

The results of the retest suggest there was greater variability than there had been in the first testing. While the freedom group had received significantly higher MTAI scores than any other group during

TABLE 2
t RATIOS FOR THE HYPOTHECATED INSTRUCTOR ATTITUDES AFTER EXPOSURE TO TEACHER ATTITUDE

Instructor Attitude	N	Mean MTAI score	*t* Ratios for differences between attitude groups								
			Patience	Learning	Clinical	Respect	No punishment	Discipline	Personal	Independence	Understanding
Freedom	20	56.00	0.95	2.39[a]	3.40[b]	1.05	1.16	6.17[c]	0	0.05	1.11
Patience	5	46.60		1.24	1.53	−0.16	−0.27	3.44[b]	−0.96	−0.65	−0.12
Learning	8	29.50			−0.11	−1.25	−1.52	1.50	−1.89	−1.62	−1.40
Clinical	6	31.00				−1.90	−2.33[a]	2.41[a]	−3.04[b]	−1.72	−2.02
Respect	12	47.77					−0.07	4.19[c]	−0.91	−0.60	0
No punishment	7	48.71						5.09[c]	−0.94	−0.54	−0.08
Discipline	16	10.00							−5.43[c]	−3.46[b]	−4.37[b]
Personal	14	55.92								0	0.99
Independence	5	55.40									0.57
Understanding	13	48.00									

Note.—Degrees of freedom are $N_1 + N_2 - 2$.
[a] Significant at P .05 level.
[b] Significant at P .01 level.
[c] Significant at P .001 level.

the pretest, the retest results indicate that the freedom group can only be distinguished from the emphasis on school learning, clinical, and the discipline groups. The emphasis on school learning is a characteristic of the authoritarian personality according to the authors of the MTAI. The results of the retest indicate further that these three authoritarian groups received significantly lower MTAI scores than many of the democratic attitude groups. The *clinical* group attained significantly lower scores than the *freedom, no punishment,* and *personal contact groups*; the *school learning* received lower scores than the *freedom* group; and the *discipline* group received significantly lower scores than all groups except the school learning. It is interesting to note that the clinical group was more democratic in attitude than the discipline group.

Again, students assigning democratic attitudes to the teacher received higher MTAI scores, indicative of a democratic attitude; students assigning antidemocratic attitudes, received lower MTAI score, indicative of an antidemocratic attitude.

SUMMARY AND CONCLUSIONS

The purpose of this study was to determine whether students' authoritarian or democratic attitudes toward children were in any way associated with their perception of their teacher.

Not having been cued to their teacher's attitude during the first phase of the experiment, the differences between those who saw their instructor as one who would encourage freedom, understand, or respect children and between those who saw him as a disciplinarian or helper of the helpless was interpreted in light of the projective hypothesis. Students, without their awareness, assigned attitudes which they themselves held.

After having been continuously cued to the teacher's attitude thereafter, the earlier vagueness disappeared experimentally, making *projection* less tenable as an explanation, but not ruling it our entirely. The writer believes the students assigning the different attitudes to their instructor operated within different frames of reference. In this study, two frames of reference were possible: a democratic and an authoritarian. Within a democratic frame of reference, students selectively perceived in the teacher evidence for freedom of, patience with, understanding of, respect of, and no punishment of children. Within an authoritarian frame of reference, students saw their instructors as being primarily interested in discipline, school learning, and children with psychological problems.

Certain attitudes which students assign to their instructor, then, differentiate between democratic and authoritarian students. In addition, there is evidence that rating of attitudes in another person will be influenced to a great degree by the attitude the rater himself holds. Hence, there is a constant error in the judgment of democratic attitudes in others. Democratic raters are apt to give more democratic ratings; authoritarian raters are apt to give more authoritarian ratings.

REFERENCES

1. Adorno, T. W., Frenkel-Brunswik, Else, Levinson, D. J., & Sanford, R. N., *The authoritarian personality*. New York: Harper, 1950
2. Chansky, N. M., How students see their teacher, *Ment. Hyg.*, 1958, **42,** 118–120.
3. Cook, W. W., Leeds, C. H., & Callis, R., *Minnesota Teacher Attitude Inventory*. New York: Psychological Corporation, 1951.

Received August 1, 1957.

Journal of Educational Psychology
Vol. 49, No. 1, 1958

THE SCHOOL PROGRESS AND ADJUSTMENT OF UNDERAGE AND OVERAGE STUDENTS[1]

CLYDE J. BAER

Kansas City, Missouri, Public Schools

The purpose of this study was to investigate the question of whether or not a child who begins school underage experiences similar problems and achieves the same level of development as if he had waited a year to enter school.

In this city school system a chronological age of five by November first is required for regular entrance into the kindergarten. However, children who will be five during November or December and who score a mental age of 5–0 or above on an individual intelligence test may be admitted. This study was made because some administrators and teachers in this school system feel that many of those children who enter prior to age five are too immature to be in school. However, there is no mid-year entrance or promotion and children who are not admitted must wait a full year to enter school.

Procedure

Seventy-three children with birthdates in November and December were matched with 73 children with birthdates in January and February who were in the same school grade and who had entered kindergarten in September of the same year. The groups were matched on the bases of intelligence quotient, sex, and, in about two thirds of the cases, the school entered.

During their eleventh year in school the groups were compared on the bases of physical size at the time of the study, grade level attained, number of problems marked on the Science Research Associates Youth Inventory and scores on the Guilford-Zimmerman Temperament Survey. From the cumulative records, comparisons were made of marks in elementary and high school subjects, achievement test scores, teacher rating on personal traits, and number of absences.

Results

Table 1 describes the groups studied.

The overage and underage students were not significantly[2] different in intelligence, as measured by the Revised Stanford-Binet, Form L, administered at time of their entrance into kindergarten or during the regular school year. For the underage students the range of IQ was from 103 to 127, and for the overage students the range was from 101 to 126.

After eleven years in school, the overage students were significantly taller but not significantly heavier than the underage students. They also had been significantly[3] more successful in maintaining regular progression from grade to grade than the underage students.

During the elementary school years (kindergarten through Grade eight) the

[1] This article is based on the writer's doctoral dissertation, the University of Kansas, 1955. Grateful acknowledgement is made to Henry P. Smith, under whose direction this work was done.

[2] Unless otherwise specified, each significant difference reported here was statistically significant at the .01 per cent level of confidence, and was obtained by the application of the t test of statistical significance, where

$$t = \frac{\bar{D}}{\sqrt{\dfrac{N\Sigma D^2 - (\Sigma D)^2}{N^2(N-1)}}}$$

[3] In this instance the chi-square test of statistical significance was applied, and indicated a level of confidence between two and five per cent.

TABLE 1
NUMBER, IQ, AGE, HEIGHT, WEIGHT, AND GRADE OF OVERAGE AND UNDERAGE STUDENTS STUDIED

	Overage		Underage	
	Boys	Girls	Boys	Girls
Number	42	31	42	31
Mean IQ	111.17	111.31	111.24	111.40
Age				
15–5			24	10
15–6			18	21
16–3	17	21		
16–4	25	10		
Height	5′7″	5′4″	5′6″	5′3″
Weight	148	120	144	115
Grade				
8			1	
9	3		5	6
10	38	30	36	25
11	1	1		

overage students were marked significantly higher than the underage students, but the differences between the overage and underage students tended to decrease as higher grade levels were reached. The girls consistently were marked higher than the boys of their respective groups, but this difference showed no identifiable tendency either to increase or decrease. In the high school the marks received by the overage students were significantly higher than the marks made by the underage students.

TABLE 2
MEAN RATINGS ON PERSONAL TRAITS FOR OVERAGE AND UNDERAGE STUDENTS

	Underage	Overage	Difference
1. Participation in Group Activity	3.61	4.38	.77*
2. Attitude Toward School Regulations	3.84	4.30	.46*
3. Appearance	4.37	4.64	.27*
4. Dependability	3.66	4.27	.61*
5. Emotional Stability	3.71	4.70	.99*
6. Initiative	3.22	4.01	.79*
7. Cooperativeness	3.64	4.28	.64*

*Significant at the .01 % level of confidence

Achievement test scores reported at various grade levels during the elementary school showed that the overage students achieved significantly higher scores in reading for Grades three, six, and eight; in arithmetic for Grades four, six, and eight and in social studies for Grade five. The difference between overage and underage were not significant in spelling for Grade five; language for Grades five and eight; and science for Grade seven.

Near the close of each school year every student is rated by his teacher on each of seven traits. Ratings recorded for each grade level from three through eight for each trait were summed and the mean taken for each pupil. Table 2 shows the mean group ratings for overage and underage students according to a scale of one to five, with five being the highest rating. For all traits the overage students were rated significantly higher than the underage students.

The girls were rated higher than the boys of their respective groups on all traits. For three traits, Attitude Toward School Regulations, Dependability, and Emotional Stability, the differences between boys and girls were greater than the differences between underage and overage.

The number of problems marked by overage and underage students on a problem inventory was not significantly different. Although the overage and underage boys tended to mark the same problems, the underage boys marked more problems relating to home and family. The overage girls marked more problems in the category dealing with school than in all the other categories combined. Although this category was also the most popular with the underage girls, they marked almost as many problems in the section "After High School."

The results on the Guilford-Zimmerman Temperament Survey indicate that the overage boys tend to be less inclined to be suspicious or to see personal reference in the words or actions of others, and the overage girls tend to be somewhat more socially aggressive, with greater drive and energy. All of the mean scores for the girls and all but five of the 20 mean scores for the boys fall in the average range.

Reported absences in elementary school show little difference between overage and underage students. From kindergarten through Grade eight the total median number of absences per year for both overage and underage boys was 6.50 days. For underage girls the total median was 8.00 days, and for overage girls it was 8.50 days. For both overage and underage the trend seemed to be for the greatest number of absences to occur in the primary grades, decrease in number during the intermediate grades, and then increase again at about Grade eight, with the girls showing the higher rate of increase.

Conclusions

As a group, the overage children made better school progress than did the underage children. The overage children, from kindergarten through Grade ten, made significantly higher marks in subjects, significantly higher scores on achievement tests in reading, arithmetic, and social studies, were rated significantly higher on personal traits by their teachers, and were significantly more successful in maintaining regular progression from grade level to grade level.

That the differences between boys and girls were greater than the differences between overage and underage for three of the personal trait ratings may indicate a sex-associated factor in these ratings. The overage and underage boys marked essentially the same problems on a youth inventory, but the two groups of girls did not show as much agreement on problems as did the boys.

Although there is some evidence that the differences between the overage and underage students tended to decrease with higher grade levels, perhaps this is what should be expected since the advantage in mental age that the overage group carries in the elementary school grades tends to decrease as the students get older.

Before concluding that it would be better for underage children to wait until the next year to begin school, it should be noted that most of the underage children made average school progress. As a group, they made average marks in subjects, average scores on achievement tests, received average ratings by their teachers on personal traits, and did not mark significantly more problems on the problem inventory than did the overage students.

However, it should be remembered that both the overage and the underage children studied here were selected on the basis of intelligence (average IQ of each group about 111). Thus, a better than average performance may legitimately be expected for either group on certain of the measures used.

Received August 4, 1957.

Journal of Educational Psychology
Vol. 49, No. 1, 1958

A FURTHER NOTE ON BASAL METABOLISM AND ACADEMIC PERFORMANCE

MARY E. YARBROUGH
Meredith College

AND HAROLD G. MCCURDY
University of North Carolina

It is sometimes necessary to conclude that an attractive hypothesis is incorrect. In 1947 one of the present authors reported in this Journal a substantial correlation between BMR and the academic performance of a sample of college women (**2**). When his correlation was combined with the lower, but still positive correlation reported in an earlier study by Patrick and Rowles (**4**), the resulting r of .24 seemed large enough to justify the speculation that college success might owe something to basal metabolic rate. It was stressed at the time, however, that the hypothesis called for a wider sampling of the college population. In the light of the further data which we can now offer, the Meredith College sample of 1947 appears as a statistical aberration and the suggested hypothesis untenable, since the true correlation between the variables at issue seems to be in the neighborhood of zero.

New and Old Data

In Table 1 are presented the data on which we base our rejection of the old hypothesis. It speaks for itself, but a few words regarding the studies which it summarizes will not be out of place.

The studies are listed in the table in roughly chronological order. It will be noticed that the information from Omwake, Dexter, and Lewis (**3**) was available in published form in 1934. It was unfortunately overlooked by McCurdy in 1947, but it would not have altered the suggestion made at that time. These authors state regarding their 72 Agnes Scott students: "Those making a high scholastic average tend to have high metabolism, but little relationship between poor scholarship and metabolism is evident." Their correlation of .138, when combined with the .05 of Patrick and Rowles and the .53 of McCurdy, contributes to a correlation (by Fisher's z transformation) of .19 for 152 individuals, which is significant at slightly better than the three per cent level.

The three new sets of data, all independently gathered, are those of Schutte at California, McCurdy at North Carolina, and Yarbrough at Meredith. The Schutte dissertation (**5**) has not been published, and the published abstract came to our attention quite recently; but the author has kindly furnished us with the important data which appear in the table in a personal communication.

Concerning the other two studies of recent date a few details can be given. McCurdy at the University of North Carolina at Chapel Hill, through the kindness of E. McG. Hedgpeth of the University Infirmary, secured a large collection of routine BMR records, of which a certain number applied to students. Some 560 of these were checked against the academic files of the Central Office of Records in order to find as many undergraduate students as possible who had taken courses falling within the central liberal arts curriculum. Many were excluded because their names did not appear in the files, or because they were graduate students, or because their curricula were irregular. The final yield was 117 cases. Point-hour ratio was ascertained for these

TABLE 1
SUMMARY OF STUDIES ON BMR AND COLLEGE SCHOLARSHIP

Study	*N*	BMR	Point-Hour Ratio	Correlation
Patrick & Rowles (Ohio University)	52 F	−7.3 ±8.97	1.50 ±.51	.05
Omwake, Dexter, & Lewis (Agnes Scott)	72 F			.138
McCurdy (Meredith)	28 F	−7.79 ±7.30	1.19 ±.55	.53
Schutte (University of California)	556 F	−13.08 ±9.31		−.06
	377 M	−12.36 ±10.62		.104
McCurdy (UNC)	102 (37 F, 65 M)	−8.50 ±10.35	1.13 ±.89	−.08
Yarbrough (Meredith)	33 F	−8.30 ±8.00	1.32 ±.72	−.05

Note.—BMR and point-hour ratio determinations are not exactly comparable from study to study because of differences in the machines used and in the assignment of points to letter grades.

cases in or near the academic quarter when the BMR was taken. These 117 students ranged in age from 17 to 34; the distribution was skewed, and it was decided to cut off the older students at a break in the distribution, leaving a total of 102 with an age range of from 17 to 24, symmetrically distributed around the mean age of 21. No significant differences appeared between the 65 men and 37 women of this sample in age, point-hour ratio, BMR, or the relations between variables, and therefore the results are lumped together in the table. It was fully realized at the time that the manner of assembling these data did not afford a clean-cut test of the hypothesis, since errors might have crept in at several points; but the specific purpose was to discover whether the hypothesized relationship was strong enough to stand up in the face of the sort of random errors which might be encountered by a practical-minded college dean or physician working upon the same hypothesis. The physician who lent the BMR records, incidentally, felt that the hypothesis was perhaps a realistic one.

Yarbrough at Meredith, on the other hand, attempted to follow strict criteria both in sampling the students and in taking the BMR readings. The students were all in their first year of college, taking virtually the same program, and being subjected to very similar requirements by teachers who value good academic work. Three BMR records were taken on each girl, and each record was carefully inspected for possible technical flaws. As in the earlier Meredith study, the record yielding the lowest BMR for a given individual was utilized in the correlational analysis. In all essential respects the experimental procedures and the nature of the sample seemed comparable to those of the earlier study. But the correlation between BMR and scholarship was definitely much lower.

It is clear from inspection of Table 1 that the correlation of .53 in 1947 has to be considered as quite exceptional. We can think of no explanation except accident of sampling. The total weight of the evidence supports the view that BMR's in the normal range have little or nothing to do with scholarship at the college level.

Conclusion and Commentary

The available evidence contradicts the plausible hypothesis that basal metabolic rate might affect academic performance in college. The correlation between the physiological and the intellectual indices used appears to be at or near zero. Perhaps the persuasiveness of the hypothesis in the first placed depended entirely too much on a superficial view of the interrelationships of energy, motivation, and thought.

In psychological circles the null hypothesis is often regarded with distaste, and "negative" results are dreaded so much that, it is rumored, some investigators avoid repeating experiments lest chance should be against them the second time. If the rumor is true, then some of our theories must have very spindly foundations. No matter what the significance level of a single result may be, it is not sufficient by itself to establish a general working rule. The present summary of metabolic studies should make it clear that one correlation in a set of six may be freakishly different from the rest.

References

1. Fisher, R. A. *Statistical methods for research workers. Edinburgh:* Oliver & Boyd, 1941.
2. McCurdy, H. G. Basal metabolism and academic performance in a sample of college women. *J. educ. Psychol.,* 1947, **38,** 363–372.
3. Omwake, K. T., Dexter, E. S., & Lewis, L. W. The interrelations of certain physiological measurements and aspects of personality. *Char. & Person.,* 1934, **3,** 64–71.
4. Patrick, J. R., & Rowles, E. Intercorrelations among metabolic rate, vital capacity, blood pressure, intelligence, scholarship, personality and other measures on university women. *J. appl. Psychol.,* 1933, **17,** 507–521.
5. Schutte, H. M. An investigation of basal metabolic rate and college scholarship. Unpublished doctoral dissertation, Univer. of California at Berkeley, 1951.

Received August 13, 1957.

Journal of Educational Psychology
Vol. 49, No. 1, 1958

AN EXPERIMENTAL PROGNOSTIC TEST FOR REMEDIAL READERS

RICHARD W. WOODCOCK
Oregon College of Education

A remedial reading program is usually a relatively expensive program since it may entail individualized or small group instruction in addition to the school's regular full-time program. Limitations of finances and time usually limit the amount of service available and it is desirable to utilize some sort of selective procedure in choosing cases for remedial instruction.

This article describes the development and evaluation of a test for selecting the remedial readers most likely to profit from special instruction. For the purposes of this discussion, a "remedial reader" is defined as any child whose reading achievement is significantly retarded below both his grade placement *and* his capacity or potential for reading achievement. In addition, the child in mind is specifically lacking in reading skills usually developed by the third or fourth grade.

The test is designed to duplicate as nearly as possible the "learning-to-read" process for remedial readers. It was assumed that the primary learning task of the remedial reader is to learn associations between written symbols and familiar spoken language. By means of the experimental test an attempt has been made to provide the remedial reader with a controlled learning situation uniquely different from reading, yet so similar to the process of learning-to-read that performance on the test might be indicative of ability to profit from remedial reading instruction.

Description of Test

The experimental test is basically composed of five short stories written entirely with pictures and symbols instead of words. Each of these "test stories" is on a successively more difficult level and is preceded by a practice page which introduces the new symbols used at that level and provides practice with the symbols in context. The manner of presenting the material on these introductory pages is similar to that of a reading lesson. Figure 1*a* shows the first group of symbols and the following line of practice material for Level I. A portion of the Level V test story is shown in Figure 1*b*.

The total vocabulary of the experimental test consists of 72 symbols. Two of these symbols represent the inflections, "-ing" and "-s". The "-s" inflection is used both as a plural and to denote possession.

The vocabulary of the first level consists of 14 symbols. These symbols represent such words as "boy," "the," "cat," "is," "black," and so forth. Many of the symbols at the first level consist of stick drawings or simple pictures of the object the symbol is intended to represent. Each successive level adds approximately 15 new symbols to the vocabulary being used in the test stories.

The symbols at the higher levels do not provide picture clues. There are some phonetic elements common to a number of the symbols. However, this is not pointed out to the *S*s.

The materials for administering the test consist of an answer sheet and a test booklet. The test booklet contains ten pages, two pages for each of the five levels, as described above. The answer sheet is used by the examiner in recording the errors made by the *S* while reading the test stories.

The score on the test is the total

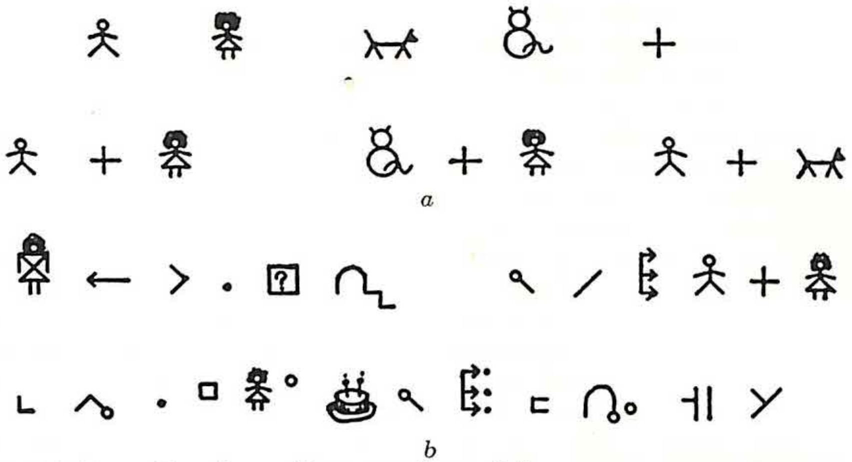

FIG. 1. (*a*) FIRST TWO LINES FROM THE LEVEL I INTRODUCTORY PAGE (Boy, Girl, Dog, Cat, And—Boy and Girl, Cat and Girl, Boy and Dog) (*b*) FIRST TWO LINES FROM THE LEVEL V TEST STORY (Mother went to a surprise party. So did the boy and girl. It was a little girl's birthday so they took presents for her.)

number of errors made on the five test stories read orally by the *S*. Hesitations of more than approximately five seconds (for which the word is supplied by the examiner), substitutions, omissions, and additions were considered errors. Repetitions were not so counted in this study.

PROCEDURE OF THE STUDY

1. The first step in the development of the instrument consisted of small-scale tryouts and revisions with a series of crude tests and procedures. From this work evolved the Experimental Prognostic Test for Remedial Readers described above.

2. The experimental test was administered to a complete fourth grade class of 24 pupils at Francis Willard School in Eugene, Oregon during March of 1955. The resulting data were analyzed for reliability and concurrent validity information.

TABLE 1
CORRELATION OF EXPERIMENTAL TEST ERROR SCORES WITH GATES READING TESTS
(N = 24)

Vocabulary	−.51
Comprehension	−.65
Speed	−.58
Oral reading	−.48
Composite score	−.59

3. Next the test was administered to 26 remedial readers in the Corvallis Public Schools for the purpose of evaluating the predictive validity. Each of these 26 remedial readers was given the experimental test and a series of Gates reading tests. Following a four-month period of daily individual and small-group instruction these same children were given an alternate form of the reading tests. These data were analyzed for predictive validity information.

FINDINGS

Concurrent validity. Twenty-four children in a regular fourth-grade classroom were administered the Gates Reading Survey, the Gates Oral Reading Test, and the Experimental Prognostic Test. The experimental test was found to correlate −.59 with the composite measure of reading achievement. (Negative correlations are based on error scores.) Table 1 presents the correlations of the experimental test with the four reading achievement subtests. These correlations are all significant at the .05 level of confidence.

A mean score of 58.0 errors was made on the experimental test by this fourth-grade class. The standard deviation of scores was found to be 22.8 errors.

Reliability. The experimental test reliability, as computed by the Kuder-Richardson Formula 21, is .91. This value probably represents an underestimate since the data do not meet the assumption of equal difficulty for all items in the test. The split-half correlation was .92 and a reliability of .96 is obtained when this value is corrected by the Spearman-Brown Formula. These values are probably overestimated to a small degree. The standard error of measurement, based on the corrected split-half reliability, is 4.6 errors. Since the standard deviation of scores, in this sample, is 22.8 errors it would seem to indicate that the test may have good discriminative properties.

Predictive validity. Estimates of the predictive value of the experimental test were obtained from a group of cases referred and selected for remedial reading instruction. Since the total number of cases was relatively small and spread among several age levels, it was not possible to apply correlational procedures to the data which had been collected. An alternative method was developed which consisted of pairing the remedial cases and predicting, on the basis of the experimental test scores, which member of each pair would show the most gain by the end of the training period. In order to pair the cases they were grouped according to age levels and each child was paired with each other child at the same age level if there was a significant difference between their scores on the Experimental Prognostic Test. Several different values were established as "significant differences" in order to observe the amount of variance in the predictive ability of the test. These significant differences in scores were arbitrarily established as two, three, five, eight, and ten times the standard error of measurement. On this basis it was usually possible to pair each child with a few other children at the same age level. The member of each pair showing the better performance on the prognostic test was predicted as the one most likely to gain from the special instruction.

This procedure experimentally duplicates a very real and practical administrative problem in the area of remedial reading. The person responsible for selecting remedial reading cases is faced constantly with the question, "Which of these two children can profit the most from remedial instruction?" In essence, this study has experimentally created a series of such possible choice situations. The statistical purpose of this portion of the study was to determine whether the proposed instrument would predict, any better than chance, which children would profit the most from instruction.

Table 2 shows the results of this portion of the study. Using two times the standard error ($2 \times s_e$) as a significant difference between prognostic test scores, it was possible to set up 44 pairs of remedial cases. Of these 44 predictions, 34 were correct—indicating a predictive efficiency of .54 (54% better than chance).

TABLE 2
PREDICTIVE EFFICIENCIES OF EXPERIMENTAL PROGNOSTIC TEST WITH CASES GROUPED BY CHRONOLOGICAL AGE
($N = 26$)

Differential (s_e units)	Pairs (n)	Correct predictions	Percentage correct	Predictive efficiency
2X	44	34	77	.54
3X	39	31	80	.60
5X	25	20½	82	.64
8X	17	15	88	.76
10X	12	11	92	.86

When higher values are used as a significant difference between scores, the number of pairs becomes less and the accuracy of prediction steadily increases. Using a value of ten times the standard error provides 12 pairs and a predictive efficiency of .86. The steady increase in predictive efficiency, as higher values are used for discrimination, is further evidence that the test is functioning predictively in this situation.

When these same cases are paired on the basis of grade level rather than age level it is possible to obtain more pairings, but there is a decrease in the predictive efficiency of the instrument. This is probably due to the more heterogenous nature of the sample in terms of age. (At one time the experimental test was administered to a number of children at various ages, and it was found that the number of errors rapidly decreases with an increase in the age of the *S*s.) If a value of five times the standard error is used as a significant difference between scores it is possible to set up 35 pairs. This yields a predictive efficiency of .60 (as compared to .64 for 25 pairs when grouped by age). A predictive efficiency of .60 was the highest obtained when pairings were made on the basis of grouping by grade (as compared to .86 when grouped by age).

Twenty-three of the cases used in this portion of the study had had Stanford-Binet Intelligence Tests administered at some time. Predictive efficiencies were computed for these data using the same procedure used to predict with the experimental test. No predictive value for the Stanford-Binet was found in this sample. However, since this portion of the study was subjected to limited control the results do need to be interpreted cautiously. One significant point is brought out by this comparison. It was quite difficult to obtain as large a number of pairings from the intelligence test data as from the experimental test data. This was due to the relatively narrow range of performance available for discrimination between *S*s on the intelligence test results as compared to the experimental test results. The range of I.Q.'s was approximately 30 points while the range of performance on the experimental test was approximately 100 points for the same *S*s. In both instances, quite coincidentally, the size of the standard error of measurement is nearly identical.

Conclusions

Within the limitations of this study it would appear that the Experimental Prognostic Test for Remedial Readers has demonstrated a predictive value in selecting cases for remedial reading instruction. Further research will be required in order definitely to determine the degree of this value. At present, it would appear that the test shows a greater possibility for *predicting* success based on gains during a period of instruction than the techniques in common use. However, it should be noted that the research concerning *the predictive value* of the techniques in current use is incomplete.

One is impressed by the paucity of research studies designed to consider some aspect of the prediction of success in remedial reading. Most studies which are referred to in this area have been originally designed for some other purpose. For example, the writing and research which provides the basis for the use of intelligence tests as a predictor of success in remedial reading is primarily based on studies of concurrent validity obtained between intelligence and reading achievement with groups of normals. These correlations have not been based on the predictive validity between the intelligence test scores of retarded readers and their subsequent gains in remedial reading. This does not mean to say that the studies were not appropriate for the purposes originally intended by the authors, but the application of the results of such studies to the problem of selecting cases or predicting success in remedial reading will remain highly questionable until subjected to experimental verification.

Received August 30, 1957.

Journal of Educational Psychology
Vol. 49, No. 1, 1958

EFFECT OF AN AUDIO-VISUAL PHONICS AID IN THE INTERMEDIATE GRADES

CAROLYN LUSER, EILEEN STANTON, AND CHARLES I. DOYLE
Loyola University, Chicago

To test the effect of formal phonics drill on a group basis, the present experiment employed audio-visual aids consisting of uniform recordings and individual charts for each pupil in four experimental rooms.[1]

The population was chosen from a lower socioeconomic area, with many semitransients. The sample contained many Negroes, Puerto Ricans, Mexicans, and a scattering of other nationalities. The area and population were selected in the hope of finding more than the average number of handicapped readers. This hope was realized when the pretests were scored. The average reading scores in all eight rooms were below the grade-level expectancy, and there were many seriously retarded readers. The number of drop-outs and absences confirmed the transient nature of the population. While nearly 300 pupils were available for pretests, only 214 completed the entire battery of tests at the conclusion of the experiment.

Two public and two parish schools were available in the area selected. The grade levels chosen for the experiment were the third and fourth grades, where reading difficulties tend to become more evident, and where pupils' competence to follow group test instructions may be assumed more safely than at lower levels.

In each school an experimental room and a control room were selected. The criterion for the selection of the experimental room was the lower average IQ derived from the Kuhlmann-Anderson Test, Form D (sixth edition). This form not only fitted the average grade expectancy of the experimental and the control rooms, but had the further advantage that the content was about equally divided between verbal and nonverbal subtests.

Besides the intelligence tests, four achievement measures were secured for each pupil; Gray's Oral Reading Paragraphs, Stanford Primary Reading, Form D (paragraph meaning and word meaning) and the Marion Monroe form (written) of the Ayres Spelling Scale (**3**).

After the pretests, each of the four experimental rooms received 43 twenty-minute sessions of phonics drill with the phonograph records and individual pupil charts (**1**). These sessions were spaced three times a week for a period of 15 weeks. The experiment was limited to 15 weeks because the pretests had to wait for mid-year promotions, and the retests had to be completed before final examinations in June. No special motivation was given during the drill periods, aside from the encouragement offered in the records themselves. All the phonics sesions were conducted by one of the workers who had

[1] Permission for the use of these four schools was graciously given by Peter B. Ritzma, Chairman, Special Projects Committee, and Don C. Rogers, Assistant Superintendent, Chicago Public Schools, and by Rev. David C. Fullmer, Assistant Superintendent, Chicago Catholic School Board. The District Superintendent, Douglas Van Bramer, and the principals and teachers in the four schools also rendered much cooperation.

The authors are indebted to the World Book Company for permission to reprint Stanford Primary Reading Test, Form D. It appeared to the authors that fill-in responses would be more valuable for detailed analysis than the multiple-choice style adopted in the 1953 forms.

The phonographs, phonics records, and pupil charts for the experimental rooms were donated by Bremner-Davis Phonics.

administered the pretests, who was thus a familiar figure to the children.

At the conclusion of the experiment, the entire population, experimental and control, was retested with the complete battery described above. Retests were conducted by the same examiners who administered the original tests.

Results

The data from both batteries were analyzed by two methods: by computation of the standard error of the gains (**2**), and by chi square. The chi-square analysis was based on the assumption that a gain in individual scores in the experimental group substantially greater than the average gain evidenced by the control group would be valid to determine a cutting point. Accordingly, for the achievement tests, a gain of .5 grade score in 3.5 months was set as significant. For the K-A IQ, allowing for practice effect much greater than that shown by the control group, a gain of three IQ points was chosen.

While the chi-square test was a less refined and precise statistic than that based on standard error, it did in general corroborate the findings of the latter. The results of both methods are summarized in Table 1.

The gains in three of the four achievement scores seem to indicate clearly the effectiveness of the standardized audio-visual drill. Failure to show a similar gain in word meaning is understandable in view of the background of the children in this sample. A marked gain in word meaning would seem to depend on enriched experience and resultant growth in vocabulary more than on mastery of phonics.

It seems evident that the unexpected gain in IQ points does not indicate a real change in intelligence. It does serve to point up the fact that performance on a group test of intelligence is markedly dependent on the pupils' familiarity with the printed word. It is suggested, however, that improved habits of attention from group drill sessions may have contributed

TABLE 1
TEST SCORES OF EXPERIMENTAL AND CONTROL GROUPS BEFORE AND AFTER AUDIO-VISUAL PHONICS DRILL

Tests	Experimental		Control		Net Diff. in gain	t	P	Chi square	P
	Pre	Post	Pre	Post					
	$N = 105$		$N = 109$						
Oral Reading	M 3.0	3.76	2.97	3.27	.46	3.65	.001	13.93	>.001
	σ 1.15	1.40	1.13	1.29					
Parag. Mean.	M 2.84	3.61	2.90	3.30	.37	4.40	>.001	8.57	.01
	σ .84	1.16	.92	.97					
Word Mean.	M 2.79	3.26	2.76	3.16	.07	1.11	.30*	1.22	.30*
	σ .78	.93	.83	.89					
Spell.	M 2.79	3.26	2.76	3.00	.23	2.35	.02	4.26	.05
	σ .97	1.17	.84	.94					
IQ	M 88.3	94.5	92.3	93.0	5.50	4.91	>.001	23.28	>.001
	σ 11.2	11.9	11.0	13.0					

* Not significant

in part to more successful performance on the K-A retests of the experimental group.

Summary

A sample of 214 third and fourth graders from four schools in an underprivileged area was divided into two groups. A control group was measured against an experimental group which received 43 drill sessions in phonics, with uniform recordings and individual charts. When both groups were retested after the experiment, the experimental group showed gains on standard tests of oral reading, paragraph meaning, and spelling, which were significantly greater than the gains of the control group. Gain in word meaning apart from context was not significant. Retest performance on a paper-and-pencil intelligence test also showed a marked gain for the experimental group.

REFERENCES

1. Bremner, A. J., & Davis, Josephine. *The sound way to easy reading.* Wilmette, Ill.: Bremner-Davis Phonics, 1953.
2. McNemar, Q. *Psychological statistics,* New York: J. Wiley, 1949.
3. Monroe, M. *Children who cannot read,* Chicago: Univer. of Chicago Press, 1932.

Received September 19, 1957.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 1, 1958

PREDICTING SCHOOL ACHIEVEMENT FOR BILINGUAL PUPILS[1]

JAMES G. COOPER

Territorial College, Agana, Guam, Marianas Islands

Teachers and other school personnel of the Territory of Guam have long been aware of the need for adequate predictors of school achievement for their bilingual pupils. In December of 1956, Guam's Department of Education authorized a study to discover to what extent, if any, currently available tests would provide such predictions. Guam is a territory whose cultural patterns are rapidly changing. From a Spanish type of church-dominated society, from a military paternalism, Guam is becoming infused with current American ideas, trends, and practices. The local language, Chamorros, is slowly giving way to English. However, English is spoken only in the school classroom, infrequently on the playground, and rarely in the home and community. Consequently, because of the unique cultural and language factors, currently available measures of intelligence must lie in the realm of questionable validity until demonstrated otherwise.

This study endeavored to determine the predictive ability of six tests of intelligence for certain fifth-grade pupils of Guam. Only those tests which were wholly or partially performance or nonverbal were considered. In order to hold cultural factors constant, four relatively isolated communities were selected. The villages of Inarajan, Merizo, Talofofo, and Umatac have had no electricity (and consequently no television) until quite recently; no statesiders (persons whose usual abode is mainland, U.S.A.) as residents; no telephones; few movies. Books and magazines are not commonly found. These four villages enrolled a total of approximately 180 fifth-grade children distributed among six classes.

The plan of study entailed the following:

1. Administer three group tests of intelligence to all fifth-grade pupils of the four villages. The tests were the California Tests of Mental Maturity, 1950 S-Form, Elementary; Davis-Eells Games, Intermediate Level; and the Culture Free Intelligence Test, Scale 2, Form A.
2. Select a stratified, random sample of 51 pupils from the larger group. Give each of these pupils the following *individual* tests: Leiter International Performance Scale, Wechsler Intelligence Scale for Children, and the Columbia Mental Maturity Scale.
3. Give all pupils the California Achievement Tests, Form AA, Elementary level.
4. Obtain teacher ratings for each child regarding his school success.
5. Relate intelligence tests scores to achievement test scores and to teachers' ratings (applicable to group tests only).
6. Study the interrelationships between certain tests.

PROCEDURE

The plan indicated above was followed: The Davis-Eells Games and California Tests of Mental Maturity were administered between November 1956 and February 1957. The Culture Free Intelligence Test was given during March of 1957 and the California Achievement Tests given in May 1957. The teachers' ratings were obtained prior to achievement testing. Individual tests were given from February through June 1957 in this sequence: Wechsler Intelligence Scale for

[1] A research supported by a grant from the James McKeen Cattell Fund.

Children, Columbia Mental Maturity Scale, and Leiter International Performance Scale.

The sample for individual testing was drawn via a table of random numbers. Each of the six classrooms was permitted to contribute its proper share of boys and girls. This was necessary because the schools of Inarajan and Merizo divided their fifth-grade pupils into fast and slow groups. Also, these schools enrolled twice as many fifth graders as Umatac and Talofofo.

Group Tests

The principal findings from the three group tests of mental ability are shown in Table 1. The table shows that 164 pupils obtained a mean California Test of Mental Maturity total IQ of 83.494 with a standard deviation of 11.087; the Pearson correlation coefficient between these scores and the California Achievement Tests was .644 and the stability of this coefficient is indicated by the .99 confidence limits of .509–.747. These limits were computed via the z transformation (**6,** p. 147). Similar data are recorded for California Language IQ, Nonlanguage IQ, Davis-Eells Games, and Culture Free Intelligence Test. The latter test is reported in raw scores because the IQs yielded a distribution severely skewed, i.e., too many low scores. The use of raw scores gave a normal (by eye) distribution.

A positive skew was noticed in the distribution of California Nonlanguage IQs. However, a chi-square test indicated that the obtained frequency distribution fitted the normal form adequately (**3,** pp. 284–285).

The Davis-Eells Games scores were considerably lower than the other two tests. The possibility existed that this may have been caused by the pupils' difficulties in understanding the directions. (This test consists of a number of pictures about which the examiner makes various statements. The subject responds by indicating which statement best applies to each picture.) This possibility was explored by retesting a fifth-grade class in Merizo. The second test was given

TABLE 1

CORRELATION COEFFICIENTS BETWEEN GROUP TESTS OF INTELLIGENCE AND CALIFORNIA ACHIEVEMENT TEST RAW SCORES

Group Test	M	σ	r	.99 confidence limits (via z)	N
California Test of Mental Maturity					
Total IQ	83.494	11.087	.644	.509–.747	164
Language IQ	81.024	10.194	.584	.434–.703	164
Nonlanguage IQ	88.067	17.239	.522	.359–.655	164
Davis-Eells Games (IPSA)	66.970	11.146	.531	.369–.601	164
Culture Free Intelligence Test (Raw scores) (this yields a mean IQ between 75 and 78)	20.685	6.355	.549	.391–.676	163
California Achievement Test					
Total raw score	154.451	36.178			164

wholly in Chamorros. The average increase of three points lacked statistical significance. It was concluded that this test measures as well in English as it does in Chamorros.

The matter of sex differences was analyzed for each of the group tests. None of the small differences approached the 5% level of significance. The mean CA of the boys was 12–0 and of the girls 11–8; the difference of four months was significant at 1%.

Discussion of group tests. The data of Table 1 indicate that the California total IQ predicted California Achievement Test scores fairly well; the schools of Guam should consider seriously more widespread application of this test. It was interesting to note that neither the Language nor Nonlanguage IQ's should be used separately.

Neither the Davis-Eells Games nor the Culture Free Intelligence Test offer as much promise. However, from the point of view of test theory, the fact that these two tests do show pronounced positive correlations with scores on a typical school achievement test is highly significant. In other words, the abilities sampled by these two measures are those which in part determine school success in a bilingual setting.

Group tests and teachers' ratings. The six teachers of the children tested were asked to rate their pupils according to the directions: "... Divide your pupils into three groups. Place the names of your best pupils into the top third, the poorest into the low third, and the others into the middle third." The relationship between this sort and the four group tests was established by locating the median for each class and preparing a 3 x 2 table from which chi square could be computed. A high value of chi square indicates that the teachers' judgments and the pupils' test positions (above or below the medians for their classes) were congruent. These data appear in Table 2.

TABLE 2
CHI-SQUARE VALUES RESULTING FROM COMPARING TEACHERS' RATINGS WITH TESTS OF ABILITY AND ACHIEVEMENT
($df = 2$)

School	*N*	Name of Test					Calif. Achievement
		Calif. Mental Maturity			Davis Eells	Culture Free	
		Total	Language	Non-language			
Inarajan							
Teacher A	29	9.48**	1.73	11.59**	2.82	5.93	19.08**
Teacher B	28	7.42*	4.47	6.36*	2.74	2.84	10.11**
Merizo							
Teacher A	23	1.30	1.92	1.81	1.60	4.71	6.19*
Teacher B	32	1.28	4.70	3.61	8.06*	2.44	16.02**
Talofofo	34	4.49	7.63*	3.52	4.77	1.07	12.77**
Umatac	26	7.26*	3.22	9.96**	1.96	10.95**	17.91**

* Significant at 5% level.

** Significant at 1% level.

The table shows that the California Test of Mental Maturity IQ's agreed fairly well with the ratings of three teachers and that the Nonlanguage IQ was as effective as the Total IQ. This finding was surprising as teachers may be expected to give more weight to their pupils' verbal behavior than to their nonverbal behavior. Neither the Culture Free Intelligence Test nor the Davis-Eells Games corresponded well with teachers' ratings. The California Achievement Test, however, showed a pronounced agreement with teachers' ratings. These latter data demonstate a high degree of validity for this test, i.e., the California Achievement Test seems to measure the kinds of achievements that these teachers deem important when they differentiate between successful and less successful pupils. These data also suggest that this test may possess considerable curricular validity. This point, however, should be verified by study of both teachers' opinions and actual curricular materials used in these classes.

Individual Tests

The results obtained from administering the three individual tests of intelligence and correlating these scores with the total raw scores from the California Achievement Tests appear in Table 3. The table shows that the best prediction was given by the Verbal Scale of the Wechsler Intelligence Scale for Children. The greatest amount of scatter was given by the Columbia Mental Maturity Scale as shown by its sigma of 24.37; this test also produced the highest mean IQ.

The means of these individual tests were well below those given in the respective manuals. The obtained sigmas were three to four IQ points lower than those reported for the Wechsler Intelligence Scale for Children (**8**), equal to the Columbia Mental Maturity Scale's listed 25 IQ points (**1**), and lower than the figures given for the Leiter International Performance Scale (**4**). The latter was difficult to determine, because the Leiter manual lacks specificity on this point. These data indicate that the IQ scores for the Guam sample were distributed in a manner similar to the standardization groups.

Discussion of individual tests. The data revealed in Table 3 show that the Verbal Scale of the Wechsler Intelligence Scale for Children gives a fairly accurate pre-

TABLE 3
PEARSON CORRELATION COEFFICIENTS BETWEEN CALIFORNIA ACHIEVEMENT TESTS, (RAW SCORES) AND IQS FROM INDIVIDUAL TESTS
(N = 51)

Test	M	σ	r	.99 confidence limits (via z)
Wechsler Intelligence Scale for Children				
Full Scale	72.89	11.84	.77	.58–.88
Verbal Scale	71.58	11.30	.80	.63–.90
Performance Scale	77.15	12.70	.54	.24–.75
Columbia Mental Maturity Scale	83.86	24.37	.61	.34–.79
Leiter International Performance Scale	72.78	12.58	.66	.40–.82
California Achievement Test (raw score, sum of all tests)	155.39	43.16		

diction of school achievement in Guam's bilingual setting. It was interesting to note that the Wechsler Full Scale IQ was not quite so efficient a predictor as was the Verbal Scale IQ. This finding was not supported by comparable evidence concerning the California Test of Mental Maturity reported in Table 1; in the latter case, the Language IQ and the Nonlanguage IQ predicted about equally well. Wechsler states that of the Verbal Scale subtests, "Information" and "Comprehension" are poor for those with inadequate verbal facility and that "Arithmetic" is influenced by education (**7**, pp. 80–82). He also notes that "Vocabulary" is affected by schooling (**7**, pp. 98–99). Therefore, it seems reasonable to believe that in a bilingual setting, successful school achievers will obtain higher scores on both the Wechsler Verbal Scale and on the California Achievement Test, i.e., we may be dealing with common elements rather than with intelligence per se.

The data relating to the Leiter International Performance Scale are of considerable concern. This test is expensive (about $200), time-consuming to administer (from 45 to 90 minutes), and very bulky. Its correlation with the California Achievement Test, although high (r equaled .66), was only slightly higher than the relatively inexpensive Columbia Mental Maturity Scale. Further, the Columbia may be administered in from 10 to 15 minutes, a factor of real importance to many psychometrists. On the positive side, however, both the Leiter and the Columbia have demonstrated considerable validity for use with these bilingual pupils. These data are all the more significant when it is recalled that the Leiter is a completely nonverbal test and the Columbia is almost so (five cards on the Columbia require the ability to read words or letters, and five involve numerals).

Interrelationships between tests. The Leiter International Performance Scale and the Columbia Mental Maturity Scale are relatively new arrivals to the scene

TABLE 4

INTERCORRELATIONS BETWEEN TESTS

(N = 51)

Test	Leiter International Performance Scale	Columbia Mental Maturity Scale
California Test of Mental Maturity		
Total IQ	.68	.62
Language IQ	.62	.54
Nonlanguage IQ	.66	.60
Wechsler Intelligence Scale for Children		
Full Scale IQ	.83	.74
Verbal Scale IQ	.73	.66
Performance Scale IQ	.78	.68
Davis-Eells I.P.S.A.[a]	.72	.69
Culture Free Intelligence Test, Raw scores (N = 50)	.75	.60
Columbia Mental Maturity Scale IQ	.69	

[a] Index of Problem Solving Ability.

of mental measurements. In order to clarify the nature of their functioning, their scores were correlated with each other and with the other tests of mental ability utilized in this study. The results are shown in Table 4. The table indicates that the Leiter measures quite consistantly with the other tests, since the correlation coefficients ranged from .62 (California Test of Mental Maturity, Language IQ) to .83 (Wechsler Full Scale IQ). The Columbia followed a pattern similar to the Leiter, but the coefficients were somewhat lower.

Summary and Conclusions

This study was undertaken to ascertain to what degree, if any, currently available measures of intelligence predict school achievement for the bilingual pupils in the Territory of Guam. Three group tests, the California Test of Mental Maturity, 1950 S-Form, Elementary; the Davis-Eells Games, Intermediate Level; and the Culture Free Intelligence Test, Scale 2, Form A were given to 164 pupils in grade five. Three individual tests of intelligence: the Leiter International Performance Scale, the Wechsler Intelligence Scale for Children, and the Columbia Mental Maturity Scale were given to a stratified, random sample of 51 pupils. School achievement was defined primarily by scores received on the California Achievement Tests, Form AA, Elementary Level, and secondarily by teachers' ratings.

All the intelligence tests correlated positively with the California Achievement Tests. The correlation coefficients ranged from .53 to .77 as follows: Davis-Eells Games, .53; Culture Free Intelligence Test, .55; Columbia Mental Maturity Scale, .61; California Tests of Mental Maturity, .64; Leiter International Performance Scale, .66; and the Wechsler Intelligence Scale for Children, Full Scale, .77.

Although teachers' ratings corresponded well with rank on the achievement test, they were not closely related to scores on the group intelligence tests.

This study demonstrated that the six intelligence tests examined predicted school success with a degree of accuracy ranging from moderate to high for Guam's bilingual pupils.

REFERENCES

1. Burgemeister, B. B., Blum, L. H., & Lorge, I. *Manual, Columbia Mental Maturity Scale.* Yonkers-on-Hudson, N. Y.: World, 1954.
2. Cattell, R. B., & Cattell, A. K. S. *Handbook for the Culture Free Intelligence Test, Scale 2.* Champaign, Ill.: Institute for Personality and Ability Testing, no date.
3. Guilford, J. P. *Fundamental Statistics in psychology and education,* (2nd ed.) N. Y.: McGraw-Hill, 1950.
4. Leiter, R. G. *Part I of the manual for the 1948 revision of the Leiter International Performance Scale.* Chicago: C. H. Stoelting, no date.
5. Lindquist, E. F. *Statistical analysis in educational research.* N. Y.: Houghton Mifflin, 1940.
6. McNemar, Q. *Psychological statistics,* (2nd ed.) N. Y.: Wiley, 1955.
7. Wechsler, D. *The measurement of adult intelligence.* Baltimore: Williams & Wilkins, 1944.
8. Wechsler, D. *Wechsler Intelligence Scale for Children; Manual.* N. Y.: Psychological Corp., 1949.

Received October 1, 1957.

Journal of Educational Psychology
Vol. 49, No. 1, 1958

THE EFFECT OF CHILD PSYCHOLOGY ON ATTITUDES TOWARD PARENT-CHILD RELATIONSHIPS[1]

FRANK COSTIN
University of Illinois

While most psychologists who teach in our colleges and universities are very much interested in changing their students' attitudes, few of them are reporting studies of the problem. In their review of research in the teaching of psychology, Birney and McKeachie (**1**) show the paucity of this kind of investigation. Furthermore, most of the attitude studies which they describe deal with the introductory course. Little is being done to discover the nature of attitudes related to advanced courses in psychology, either on the graduate or undergraduate level. The study presented in this paper is intended to help close that gap.

A second-level course which is becoming increasingly popular with undergraduates is child psychology. The students who take this course vary considerably in their educational and vocational goals, but almost all of them share at least one common objective: they hope to learn something about children which will help them as future parents. They are especially interested in learning how their own attitudes might affect their children's development. It is important, therefore, that those who teach this course find out what kinds of attitudes toward parent-child relationships their students have, and how these attitudes change when they study child psychology. As a step in this direction, a recent pilot study by Hurley and Laffey (**2**), involving 19 students, concluded that a 10 week's child psychology course at Michigan State University succeeded in making these students less rejecting in their attitudes toward children. There was no change in overprotecting attitudes.

The present paper reports a more extensive investigation of the same kind of problem explored by Hurley and Laffey. It attempts to answer this question: To what extent can an undergraduate, one-semester course in child psychology change students' attitudes toward parent-child relationships?

Method

The subjects of this study consisted of four different classes totaling 157 students. None of them had previously taken a course in either child psychology, child development, or family relations. They represented a variety of majors within the College of Liberal Arts and Sciences at the University of Illinois, and were also drawn from other colleges within the university.

The content of the course was the same for each class, with emphasis on problems of parent-child interaction. All classes were taught by the same instructor. In addition to the use of a basic text and lectures as sources of information and attitudes, films were shown and discussed. Supplementary readings also helped to extend the variety of course content.

A parent-child attitude scale was administered to each class at the beginning and end of the course. The scale was also given at the beginning and end of the semester to 155 undergraduates at the same university who were enrolled in a one semester introductory course in sociology. This was done to see whether or not changes which might occur in the psychol-

[1] A briefer version of this paper was read on September 4, 1957, at the convention of the American Psychological Association in New York City.

ogy students could be simply the result of general college living rather than taking child psychology. Like the psychology classes, the sociology students represented a variety of majors within the College of Liberal Arts as well as other colleges. They were approximately the same as the psychology students in age and college class. None had ever taken a course in child psychology, child development, or family relations, nor was any taking one at the times the scale was administered.

The attitude scale was a slightly modified version of one constructed and validated by E. J. Shoben (**4**). Ten items from Shoben's scale were omitted, because they represented a miscellaneous group which lacked homogeneity, and therefore were not as meaningful for this study as the other items. In all other respects the scale was identical with Shoben's. It consisted of 75 statements about parent-child relationships to which students responded by strongly agreeing, mildly agreeing, mildly disagreeing, or strongly disagreeing. The scale measured three kinds of attitudes: *dominating* (40 items), *possessive* (20 items), and *ignoring* (15 items).

Dominating attitudes "reflect the tendency on the part of the parent to put the child in a subordinate role, to take him into account quite fully but always as one who should conform completely to parental wishes under penalty of severe punishment" (**4**, p. 137). The following statements from the scale represent this kind of attitude: "A child should have strict discipline in order to develop a fine, strong character." "It is sometimes necessary for the parent to break the child's will."

Possessive attitudes "reflect the tendency to 'baby' the child, to emphasize unduly (from a mental hygiene viewpoint) the affectional bonds between parent and child, to value highly the child's dependence on the parent, and to restrict the child's activities to those which can be carried on in his family group" (**4**, p. 137). Statements illustrating this attitude are: "The best child is one who shows lots of affection for his mother." "Children should always be loyal to their parents above anyone else."

Ignoring attitudes reflect "the tendency on the part of the parent to disregard the child as an individual member of the family, to regard the 'good' child as one who demands the least parental time, and to disclaim responsibility for the child's behavior" (**4**, p. 137). Sample statements representing this kind of attitude are the following: "Parents cannot help it if their children are naughty." "Quiet children are much nicer than little chatter-boxes."

All scoring of responses followed Shoben's weighting system, a procedure he had found to result in parents of problem children making significantly higher scores, on the average, than parents of nonproblem children. Scores were interpreted as follows: The higher the score, the more intense is the attitude; the lower the score, the less intense is the attitude.

Results

As Table 1 indicates, the psychology students expressed a significant decrease in the intensity of their attitudes, as measured by the total scale. The mean change was 7.63 scale points. This shows that in general the attitudes of the psychology students toward parent-child relationships became more permissive by the end of the course.

The psychology classes also changed significantly in each of the three attitude dimensions which made up the total scale. Table 1 shows that the greatest change was in dominating attitudes, with a mean decrease of 5 scale points. Changes in possessive and ignoring attitudes were smaller, being 1.51 and 1.12 scale points respectively.

These changes were evidently the result of taking child psychology, since an analysis of the sociology classes showed no significant change in attitudes. Table 2 reveals this lack of change between the pretest and posttest scores. A comparison of the initial attitudes of the sociology and psychology students shows that both groups expressed essentially the same intensity of attitudes at the beginning of their respective courses. Only the psychology students, however, changed significantly. This fact supports the conclusion that the psychology students changed as

TABLE 1

ATTITUDES TOWARD PARENT-CHILD RELATIONSHIPS BEFORE AND AFTER COURSE IN CHILD PSYCHOLOGY

($N = 157$)

Attitude measured	Before course		After course		Difference in means	t
	Mean score	*SD*	Mean score	*SD*		
Dominating	154.55	11.23	149.55	10.18	5.00	7.46*
Possessive	73.50	6.30	71.99	5.82	1.51	3.75*
Ignoring	52.40	3.87	51.28	3.86	1.12	3.29*
Total scale	280.45	16.30	272.82	15.93	7.63	7.00*

Note.—The higher the score, the more intense is the attitude.

* $p < .01$.

TABLE 2

ATTITUDES TOWARD PARENT-CHILD RELATIONSHIPS OF STUDENTS WHO DID NOT TAKE COURSE IN CHILD PSYCHOLOGY

($N = 155$)

Attitude measured	Before course		After course		Difference in means	t
	Mean score	*SD*	Mean score	*SD*		
Dominating	154.36	11.12	154.24	12.29	.12	.17
Possessive	74.44	5.66	74.21	5.04	.23	.56
Ignoring	51.41	4.45	51.58	4.27	.17	.45
Total scale	280.05	16.23	280.21	16.74	.16	.14

Note.—The higher the score, the more intense is the attitude.

TABLE 3

ATTITUDES TOWARD PARENT-CHILD RELATIONSHIPS OF STUDENTS ACHIEVING IN UPPER AND LOWER HALVES OF COURSE IN CHILD PSYCHOLOGY

	Before course		After course		Difference in means	t
	Mean score	*SD*	Mean score	*SD*		
Upper half ($N = 79$)	275.84	14.56	268.54	14.46	7.30	5.84*
Lower half ($N = 78$)	285.65	16.11	277.29	15.93	8.36	5.61*

Note.—The higher the score, the more intense is the attitude.

* $p < .01$.

TABLE 4

Men and Women Students' Attitudes Toward Parent-Child Relationships Before and After Course in Child Psychology

	Before course (total scale)		After course (total scale)		Difference in means	*t*
	Mean score	*SD*	Mean score	*SD*		
Men ($N = 65$)	281.05	15.83	272.79	15.73	8.26	5.66*
Women ($N = 92$)	280.14	16.50	273.02	15.91	7.12	4.98*

Note.—The higher the score, the more intense is the attitude.
* $p < .01$.

a result of taking a course in child psychology, and not simply because of general college experiences.

In order to discover what relationship might exist between scholastic achievement in child psychology and changes in attitudes, a comparison was made between students achieving in the upper half of their class and those achieving in the lower half. The basis for comparison was scores obtained on objective examinations covering course content, including a comprehensive final examination. Results of this analysis are summarized in Table 3. Initially the upper half of the class was more permissive than the lower half. Both groups changed significantly. The amount of change, however, for each group was the same. (The actual difference of 1.06 has a *t* value of .59.)

Did men and women students differ in their attitude changes? The answer to this question can be found in Table 4. Initially, the attitudes of both groups were approximately the same. Both men and women showed a significant decrease in attitude intensity. The amount of change, however, for each group was the same. (The actual difference of 1.14 has a *t* value of .59.)

Discussion

At the conclusion of a course in child psychology, students expressed more permissive attitudes toward parent-child relationships than they had held at the beginning of the semester. While they changed significantly in all three kinds of attitudes measured, their dominating attitudes decreased the most. Why did the greatest change occur in that particular area? There are several factors which should be considered in answering this question. All of them, to some extent, probably played a part in effecting change.

First, it may be that for the kind of population which made up the psychology classes, dominating attitudes are more susceptible to change through formal education than are possessive and ignoring attitudes. Most of these students were not very far removed in time and memory from their adolescence. Since problems of parental domination are so outstanding in typical adolescent-parent relationships, it may be that recent sensitization made these students more keenly aware of such problems. Thus they would be ready to react to any course content which depicted the effects overdominating parents have on children.

A second important factor is that the students may simply have examined more course content dealing with dominating attitudes, and thus were able to change more. Of course, this possibility is closely related to the first. They may have encountered more material dealing with parental domination because they were more ready to react to it, and therefore sought it out more often than information dealing with possessive and ignoring attitudes. (The instructor of the course is satisfied that neither his lectures nor assignments were more heavily weighted with material dealing with dominating attitudes than with other kinds of parent-child relationships.)

A third likely reason for the greatest change being in dominating attitudes lies in the nature of the attitude scale. It will be recalled that the dominating area contained 40 items, as compared with 20 for possessive and 15 for ignoring attitudes. Thus, the dominating area of the scale afforded a wider range of possible responses, and therefore more opportunity to express change. Furthermore, this part of the scale had a higher correlation with the total scale (.86) than did the other two attitude categories (**4**, p. 131). It is also significant to note that when Shoben compared the attitudes of parents of nonproblem children with parents of nonproblem children, he found the greatest difference in the two groups to be in dominating attitudes (**4**, p. 132).

The fact that low achievers changed as much in their attitudes as high achievers confirms something teachers have strongly suspected for a long time: Grades don't tell them all they would like to know about what students obtain from a course! However, had measurements of scholastic progress other than objective examinations been used, the relationship between change in attitude and achievement might have been different.

Evidently the course had the same effect on the attitudes of men and women students, both groups becoming more permissive to the same extent. This is a desirable outcome, since it is a commonplace that both parents play important parts in affecting their children's development.

Considering the great amount of deliberate emphasis on parent-child relationships built into the course, the investigator had anticipated greater changes in attitudes than those discovered. But perhaps it is unrealistic to expect in one semester much more change than this. Then, too, these measured changes may be only the initial stimulation for greater changes which will take place later in the students' development.

In any case, however, studies like this one would probably be more meaningful if they measured more specific attitudes than those represented by Shoben's threefold classification. The investigator is now in the process of doing this by using an inventory recently developed by Schaefer and Bell (**3**). Their instrument is an improvement over Shoben's in several respects, an important one being that its attitude areas, delineated on the basis of factor analysis, are more specific and homogeneous than the areas used by Shoben. It is hoped that others interested in measuring attitudinal outcomes of child psychology will also try out this instrument.

Summary

Will an undergraduate course in child psychology change attitudes toward parent-child relationships? To answer this question, an attitude scale was administered to 157 students before and after the course. Postcourse attitudes were significantly less dominating, possessive, and ignoring. The greatest change was in dominating attitudes. Scholastic achievement in the course was not significantly related to the amount of change. Men and women did not differ significantly in how much they changed. As a control measure, the same scale was administered to 155 students before and after a course in introductory sociology. No change in attitudes occurred.

REFERENCES

1. Birney, R., & McKeachie, W. The teaching of psychology: A survey of

research since 1942. *Psychol. Bull.,* 1955, **52,** 51–68.

2. Hurley, R., & Laffey, J. Influence of a conventional child psychology course upon attitudes toward children. *Papers of the Mich. Acad. Sci., Arts, and Letters,* 1957, (1956 meeting), 299–306.
3. Schaefer, S., & Bell, R. Q. Parental attitude research instrument (PARI). Bethesda, Md.: Nat. Inst. of Mental Health, Child Development Section, no date (Mimeo.).
4. Shoben, E. J., Jr. The assessment of parental attitudes in relation to child adjustment. *Genet. Psychol. Monog.,* 1949, **39,** 101–148.

Received October 7, 1957.

Journal of Educational Psychology
Vol. 49, No. 1, 1958

THE DIFFERENTIAL EFFECT OF MANIFEST ANXIETY ON TEST PERFORMANCE[1]

EVAN W. PICKREL

Air Force Personnel and Training Research Center

Though the Taylor-Spence anxiety-drive concept has received considerable attention in the experimental literature, there is a significant lack of information pertaining to the effect of this variable on psychometric performance. The thesis of this study, in accordance with evidence available in the experimental literature, is that manifest anxiety level may be shown to have a differential effect on various phychometric measures; and further, that this differential effect may be shown to vary with the number of alternatives available for decision-making, the amount of habit interference possible within the task. The anxiety-drive hypothesis essentially states that the higher the anxiety score on the manifest anxiety scale (**6**), the stronger the excitatory potential and the greater the response strength. For example, a group scoring high on the scale has been demonstrated (**3**, **4**, **5**) to be superior to a low scoring group in the amount of conditioning exhibited. In more complex learning situations, however, in which there are a number of alternative competing tendencies, it has been shown (**1**) that the effect of increasing a drive would depend upon the initial response hierarchy and the relative habit strength of the correct or goal attaining response in the hierarchy. Differences in drive level may lead to superior performance by either the anxious or the nonanxious group depending on the difficulty of the choice points.

It is therefore assumed that there would be no significant difference in the performance of anxious and nonanxious *S*s on simple speeded measures which pose problems not having alternative competing tendencies. However, when the test problems offer a minimal number of alternative competing tendencies and thus are parallel to the experimentalist's conditioning problem in terms of possible habit interference, the performance of anxious *S*s will be superior to that of nonanxious *S*s. Furthermore, when the measures become quite complex and each problem offers a number of alternative competing tendencies, the performance of nonanxious *S*s will be superior to that of anxious *S*s.

Method

Measures

Four measures were selected to test the hypotheses. Answer Sheet Marking Test B1710AX was chosen as a simple repetitive task which does not bring forth any alternative competing choices for solution of the problems. The test is used to determine how quickly and accurately *S*s can locate and mark designated responses on a 15-choice IBM answer sheet. There are two parts, each presenting 75 randomly distributed items.

Army Clerical Speed ACS-2 was selected as a measure in which the test problems offer a minimal number of alternative competing tendencies. The test presents 125 four- to seven-digit numbers paired with either their reversals or a reasonable approximation. The task is to state

[1] This report is based on work done under ARDC Project No. 7719, Task No. 17011, in support of the Research and Development Program of the Air Force Personnel and Training Research Center, Lackland Air Force Base, Texas. Permission is granted for reproduction, translation, publication, use, and disposal in whole and in part by or for the United States Government.

whether an exact reversal of the stimulus number is presented.

A Code Learning Test was developed to parallel the complex learning situations often used in experimental laboratories. The test utilizes a tape recorded and IBM answer sheets for group administration, and teaches the first 10 letters of International Morse Code by presenting each signal twice, testing at the end of the sequence, and repeating the procedure 10 times. Part scores were obtained for each of the 20-item subtests. All *S*s reporting previous experience with the code were eliminated from the experiment.

A simple, nonspeeded arithmetic achievement test was selected as a complex measure of overlearning. California Achievement Test II, a subtest 4, section D of parallel forms AA, BB, and CC were used. Each form presents 20 addition problems varying in complexity from the summation of: 2 two-digit to 4 four-digit whole numbers, decimals, fractions and complex numbers when presented in columnal and linear array. Scores were obtained for each parallel form of the test.

Subjects

All *S*s were newly inducted basic airmen entering Lackland Air Force Base from November 29, 1955 to January 20, 1956. A total of 723 *S*s were administered a modified form of the Taylor scale which included the 50 anxiety items of the original scale plus the Lie and Manic scales from the Minnesota Multiphasic Personality Inventory. The Lie scale was used to detect false scores, a score of eight or above eliminating an *S* from selection. The anxious and nonanxious groups each consisted of those whose scores fell respectively in the upper and lower 20% of scores for a standardization population of about 1000 basic airmen. The cutoffs were 21 and higher, 10 and lower respectively on the manifest anxiety scale. There were 178 *S*s in the anxious group and 159 *S*s in the nonanxious group.

Anxious and nonanxious *S*s from each flight of approximately 60 men were presented the above measures in the order described.

Results and Discussion

The means and standard deviations for anxious and nonanxious groups on the Answer Sheet Marking and Army Clerical Tests are presented in Table 1. The analysis of variance technique was used to test for significant differences between performances of the two groups on the complex learning and achievement measures. This information is presented in Table 2.

In all cases the hypotheses were substantiated. There was no significant difference in the speed and accuracy with which anxious and nonanxious *S*s could locate and blacken specified spaces on an IBM answer sheet. The mean performance score for anxious *S*s was significantly higher than that of nonanxious *S*s on the Army Clerical Speed Test. On the more complex activities, the Code Learning Test and the arithmetic achievement tests, the mean performance scores of nonanxious *S*s were significantly higher than those of anxious *S*s.

Performance on the Army Clerical Speed Test parallels the results obtained with conditioning problems in the experimental laboratories. On the more complex activities, the code learning and arithmetic achievement tests, performance parallels the results obtained with more complex learning situations in the experimental laboratories. Since these results parallel those found with conditioning and complex learning situations, the appropriate placement of the control task on a complexity continuum is open to question. The amount of habit interference possible within this task does not seem as great as within the Army Clerical Speed Test. Yet perform-

TABLE 1
Mean Performance Scores

Measure	Groups				*t*
	Anxious ($N = 178$)		Nonanxious ($N = 159$)		
	Mn	*SD*	Mn	*SD*	
Answer Sheet Marking	109.51	23.92	112.04	23.24	.98
Army Clerical Speed	70.70	27.75	63.34	26.47	2.48*

* Significant at the .02 level.

TABLE 2
Analysis of Variance for Testing Significant Differences Between Performance of the Two Groups

Source	Code Learning Tests		Calif. Achiev. Tests	
	df	Variance	*df*	Variance
Between groups	1	253.09*	1	114.08*
Between trials	9	1396.94*	2	130.99*
Between individuals within groups	335	105.87*	335	40.02*
Groups by trial interaction	9	3.35	2	2.02
Residual	3015	5.29	670	2.96

* Significant at the .01 level.

ance on the Army Clerical Speed Test paralleled the results of conditioning experiments. Either the control task is misclassified on a complexity continuum or the Taylor Scale does not function as a measure of drive in this situation. A study has been initiated to provide the answer to this question.

Previous studies have shown that an extremely high scoring group on a manifest anxiety scale is constantly superior to a low scoring group in the amount of conditioning exhibited but that in more complex learning situations a high drive level could result in an impairment of performance. In this study a control task was also introduced, a situation in which there was no significant difference between the performance of anxious and nonanxious *S*s. Thus the differential effect of manifest anxiety on performance was shown with psychometric variables to be a function of the number of alternatives for decision-making given to *S*. Since results were obtained which are consistent with those found in experimental literature, the study has demonstrated a parallelism between psychometric measures and some types of variables studied in experimental laboratories.

Summary

Groups of anxious and nonanxious *S*s, as determined by a scale of manifest anxiety, were administered psychometric tasks which varied in the number of alternatives for decision-making given to *S*. There was no significant difference in the performance of the two groups on a control measure which did not provide *S* with alternatives for decision-making. When the task provided a minimal number of alternatives, mean performance of the anxious group was superior. When the task was complex and a number of alternatives were made available, mean performance of the non-

anxious group was superior. Results were obtained which are consistent with those found in the experimental literature, thus demonstrating a parallelism between psychometric measures and some types of variables studied in experimental laboratories.

REFERENCES

1. Farber, I. E., & Spence, K. W. Complex learning and conditioning as a function of anxiety. *J. exp. Psychol.*, 1953, **45,** 120–125.
2. Sarason, S. B., Mandler, G., & Craighill, P. C. The effect of differential instructions on anxiety and learning. *J. abnorm. soc. Psychol.*, 1952, **47,** 561–565.
3. Spence, K. W., & Taylor, Janet. Anxiety and strength of the USC as determiners of the amount of eyelid conditioning. *J. exp. Psychol.* 1951, **42,** 183–188.
4. Spence, K. W., & Taylor, Janet. The relationship of anxiety level to performance in serial learning. *J. exp. Psychol.*, 1952, **44,** 61–64.
5. Taylor, Janet. The relation of anxiety to the conditioned eyelid response. *J. exp. Psychol.*, 1951, **41,** 81–92.
6. Taylor, Janet. A personality scale of manifest anxiety. *J. abnorm. soc. Psychol.*, 1953, **48,** 285–290.

Received October 11, 1957.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 1, 1958

SOME SOCIOECONOMIC CORRELATES OF ACADEMIC APTITUDE

KHOSSROW MOHANDESSI AND PHILIP J. RUNKEL

Bureau of Educational Research, University of Illinois

It is usually agreed that the socioeconomic status of the family and community in which children and youth grow up affects their learning behavior and school achievement (e.g., **2**, **3**, **5**, **10**). Davis, for example, explains the influence of social class in these words: "By defining the group with which an individual may have intimate clique relationships, our social class system narrows his training environment. His social instigations and goals, his symbolic world and its evaluation are largely selected from the narrow culture of that class with which he can associate freely" (**2**, p. 609).

In the present study, academic aptitude of students in Illinois high schools is compared with some of the socioeconomic characteristics of various communities in which the schools are located. Variables chosen for this purpose were those for which there was ready-to-hand information and which tended to "narrow the culture" of the people in the community. They are the distance from large towns and cities, community population, distance from the nearest active coal mine, the value of farm products sold, the value of land and building per farm, the school size, and whether the school is public or private. Obviously, these are not direct measures of the psychological effects of family and community but rather are rough indices which we presume will be associated with psychological processes which in turn are related to academic aptitude. While these variables are not of the kind to please the theoretician, they have the advantage for the practical worker that the necessary information is readily available.

Although this study was made only of schools within the state of Illinois, there is evidence that the relationships reported here would be typical of those throughout the nation. In a nationwide sampling of high schools, Mollenkopf and Melville (**7**) found that academic aptitude and achievement were related to such variables as region (South or non-South), community size, percentage of fathers who were high school graduates, instructional support per pupil, percentage of support from state aid, and whether the region served by a school had a public library.[1]

Somewhat similar results were obtained by Thorndike (**8**) in his study of community variables as predictors of intelligence and academic achievement. Correlations between IQ and various indices of the education of adult population ranged from .33 to .43 and between IQ and different measures of cost of housing from .30 to .32. A correlation of .28 was obtained between IQ and proportion of native-born whites, one of −.26 between IQ and rate of female employment, and one of .28 between IQ and frequency of professional workers in community.

In the present study, the academic aptitude of the students was measured by the portion of the Differential Aptitude Tests (**1**) which is used as a part of the Illinois Statewide High School Testing Program. The score used is the Total score from the Abstract Reasoning and Verbal Reasoning Tests. This score will be referred to as the "D.A.T." in this paper. We will also talk of the Illinois Statewide Testing Program as the "Program."

From the 1955–56 Program we have figures available showing the mean D.A.T. for each school computed from the per-

[1] This prepublication memorandum is cited with the kind permission of the authors.

centile scores of the pupils in that school. Using this information, we chose as our "dependent" variable an extreme dichotomy; namely, whether the school mean lies in the top quarter of all schools or in the bottom quarter. The second and third quarters are ignored.[2] It should be remembered that not all high schools in the state appear in the Program and correspondingly in this analysis. The schools included are those participating in the Statewide Testing Program and having school means on the D.A.T. falling either in the highest or the lowest quarters.

Hypothesis

The purpose of this study was to check some rough indices of the socioeconomic characters of different communities in the state of Illinois (or of the schools which are located in those communities) against the mean academic aptitude of the students sent to schools in those communities. Our general hypothesis is that *some of the factors which determine aptitudes in schools in various communities are to be found among the socioeconomic characteristics of those communities*. We would suppose that, while some of this relationship is due to the efficacy with which certain types of social situations foster the growth of academic ability, another part of the relationship is due to the self-selection of the kinds of families which move into the various kinds of communities. Our prediction in each case below stems from the expectation that the more a community is able to support its schools, the greater are the chances that the students will show the higher ranges of aptitude.

TABLE 1
MEAN D.A.T. VERSUS DISTANCE FROM NEAREST TOWN OF 25,000 OR LARGER
($N = 167$)

Distance of schools from towns of 25,000 or larger	Quarter of mean D.A.T. of school	
	First	Fourth
Less than 10 miles	47	12
Over 10 to 25 miles	35	20
More than 25 miles	12	41

Note.—$p < .001$ by chi square for 2 *df*.

Findings

A description of the socioeconomic measures used in the study and the results obtained in each case follow.

Distance from Town of 25,000 or Larger

First examined was the distance of each school from the nearest town of 25,000 or more in population. These distances were scaled from an ordinary current highway map. Communities farther from the larger towns, we felt, would be more rural in character. Table 1 shows the distribution of schools in the first and last quarters of the mean D.A.T. scores according to their distance from the nearest town of 25,000 or more. The mean distance from such towns is 12 miles for the schools in the first (highest) quarter of mean D.A.T. scores and 29 miles for the schools in the fourth (lowest) quarter. The chi-square test applied to the data of Table 1 gives a value significant at the .001 level and supports the prediction made.[3]

It should be noted, in regard to Table 1 and all other tables, that we chose the categories of the independent variable so

[2] Since cases falling in the second and third quarters are not examined, conclusions cannot be drawn about these quarters. We can only draw conclusions about whether a variable (such as distance from larger towns) is associated with the school's falling in the *first* or *fourth* quarter.

[3] We report in each case the p value which is the smallest the table lists for the obtained value of chi square. We report these values so that the reader may choose for himself the level of probability which he wishes to consider statistically significant.

as to keep the frequencies in each category as equal as might be while at the same time avoiding expected frequencies too small to satisfy the requirements of the chi-square test.

Community Size

A second variable used was the population of the communities in which the schools with higher and lower D.A.T. means were located. The assumption was that larger centers of population would be able to spend more money on their schools than the smaller ones. The relation was examined twice: once including schools in the Chicago metropolitan area and a second time excluding them. Population figures were taken from the highway map. Towns not listed on the map were assumed to have populations under 1,000. As shown by Tables 2(a) and 2(b), the expected relation is significant at the .001 level whether or not Chicago schools are included.

TABLE 2
MEAN D.A.T. VERSUS SIZE OF COMMUNITY

Population of the community where schools are located	Quarter of mean D.A.T. of school	
	First	Fourth
(a) Including Chicago Schools[a] ($N = 167$)		
Under 1,000	14	33
1,001–5,000	25	28
5,001–25,000	27	10
Over 25,000	28	2
(b) Excluding Chicago Schools[b] ($N = 144$)		
Under 1,000	14	33
1,001–5,000	25	28
Over 5,000	33	11

[a] $p < .001$ by chi square for 3 *df.*
[b] $p < .001$ by chi square for 2 *df.*

Distance from Nearest Active Coal Mine

The next prediction was that proximity to coal mines would tend to lower the mean aptitude to be found in the school. Table 3 shows the distribution of schools in the first and fourth D.A.T. quarters according to their distances from the nearest active coal mine. This variable was chosen because a great number of people, especially in the southern part of the state, make their living through working in the coal mines. Some of the farmers work in them to earn "extra" money. Thus, existence of coal mines becomes an important factor in shaping the economy of an area and in deciding the kind of population inhabiting certain localities, particularly in the southern half of the state. The appropriate information was taken from a map of mineral industries (4). The relation predicted is significant at the .001 level.

TABLE 3
MEAN D.A.T. VERSUS DISTANCE FROM NEAREST ACTIVE COAL MINE
($N = 167$)

Distance from nearest active coal mine	Quarter of mean D.A.T. of school	
	First	Fourth
Less than 20 miles	16	31
Over 20–35 miles	18	20
Over 35–50 miles	33	11
Over 50 miles	27	11

Note.— $p < .001$ by chi square for 3 *df.*

Value of the Farm Products Sold in the County

The value of the farm products sold in different counties of the state was used as an index of the income of the residents of a given locality. The analysis was made separately for counties containing towns of 25,000 or over and for counties not containing such towns in order to find out whether there was any difference between the two kinds of counties in this respect. It was thought that where value of farm products sold was higher, urban and rural areas would both prosper and property values would be high in general. It was expected that schools with the higher D.A.T. means would be located in communities which showed the greater values of farm products sold during the year. Values were taken from the 1954 U. S. Census of Agriculture (**9**). Tables 4(a) and 4(b) present the relevant data. The expected relation was significant for both categories, i.e., counties with and without larger towns, beyond the .05 level.

TABLE 4
MEAN D.A.T. VERSUS VALUE OF FARM PRODUCTS SOLD

Value of farm products sold in millions of dollars	Quarter of mean D.A.T. of school	
	First	Fourth
(a) Counties Containing Towns of 25,000 or Larger[a] ($N = 68$)		
0–14.99	5	7
15.0–24.99	30	4
25.0 and above	17	5
(b) Counties Not Containing Towns of 25,000 or Larger[b] ($N = 99$)		
0–14.99	15	29
15.0–24.99	12	20
25.0 and over	15	8

[a] $p < .01$ by chi square for 2 *df.*
[b] $p < .05$ by chi square for 2 *df.*

Value of Land and Building Per Farm

As a second index of income and therefore of the ability of rural people to give financial support to their schools, the average value of land and buildings per farm was used. This likewise was computed separately for counties containing towns of 25,000 or over and counties not containing such towns. The information was taken from the 1954 U. S. Census of Agriculture (**9**). Tables 5(a) and 5(b) contain the data relevant to this question. The partitioning of chi square (**6**) was used to test the significance of the relations among the variables involved; i.e., quarter of D.A.T., average value of land and buildings per farm, and whether or not the county contained any town of 25,000 and over. The only significant relation—at the .001 level—was that between quarter and location in counties with towns of 25,000 or over. This latter relation is of little interest here since it is only another form of the relation between mean D.A.T. and the distance from larger towns. In brief, our expectation was *not* borne out that the mean D.A.T. of the school would be related to the value of farm land and buildings in the county. We have no ready explanation of the reason that this variable should fail to show a significant relation while our other variables, just as imprecise, do so.

TABLE 5
MEAN D.A.T. VERSUS AVERAGE VALUE OF LAND AND BUILDINGS PER FARM

Value of land and buildings per farm in thousands of dollars	Quarter of mean D.A.T. of school	
	First	Fourth
(a) Counties Containing Towns of 25,000 or Larger ($N = 68$)		
Less than 45.0	31	8
More than 45.0	21	8
(b) Counties Not Containing Towns of 25,000 or Larger[b] ($N = 99$)		
Less than 45.0	17	32
More than 45.0	25	25

[a] $p > .10.$
[b] $p > .10.$

Size of School

Another variable used was the size of the school itself. This information is avail-

TABLE 6

MEAN D.A.T. VERSUS SCHOOL SIZE*

(N = 138)

Number of pupils in school	Quarter of mean D.A.T. in school	
	First	Fourth
1–99	5	23
100–299	28	36
300–499	13	8
500–999	11	4
More than 1,000	10	0

Note.—The information concerning this variable was lacking for 29 schools in the Program.

* $p < .001$ for chi square for 4 *df.* (computed with the Yates correction for continuity)

TABLE 7

MEAN D.A.T. VERSUS PUBLIC-PRIVATE VARIABLE

(N = 167)

Public -private variable	Quarter of mean D.A.T. of school	
	First	Fourth
Public schools	66	71
Private schools	28	2

$p < .001$ by chi square for 1 *df.*

able from the files of the Illinois Statewide High School Testing Program. The prediction was that schools in the first quarter would be significantly larger in size than the schools in the fourth quarter. Table 6 contains the data relevant to this point. The prediction is supported at the .001 level.

The Public-Private Variable

This measure was used as another variable which might correlate with mean aptitude in the school. It was predicted that the ratio of private schools to public schools would be significantly larger in the first quarter of D.A.T. scores than in the fourth quarter. Here one would suppose the effect to be due largely to the selectiveness of the private school. Table 7 presents the relevant data. The frequencies in Table 7 seem at first glance very "lopsided." The reason for this, of course, is that so many more public than private schools participate in the Illinois Statewide Testing Program. Nevertheless, the chi-square value for the relation in Table 7 is significant at the .001 level.

SUMMARY AND CONCLUSION

The purpose of the present study was to investigate some socioeconomic correlates of academic aptitude. The study was intended to be exploratory rather than precise. Academic aptitude was examined by using schools whose mean test scores were in the upper or lower quarters among schools participating in the Statewide Testing Program of the University of Illinois. Findings of the study were that the mean academic aptitude of the school was strongly related to (a) the distance from larger towns, (b) the community population, (c) the distance from active coal mines, (d) the value of farm products sold in the county where the school is located, (e) the size of the school, and (f) whether the school is public or private.

REFERENCES

1. BENNETT, G. K., SEASHORE, H. G., & WESMAN, A. G. *Differential aptitude tests,* Form A. New York: The Psychological Corp., 1947.
2. DAVIS, A. Child training and social class. In R. G. Barker, J. S. Kounin, & H. F. Wright (Eds.) *Child behavior and development.* New York: McGraw-Hill, 1943. Pp. 607–619.
3. DAVIS, A. American status systems and the socialization of the child. In C. Kluckhohn, & H. A. Murray, (Eds.) *Personality in nature, society, and culture.* New York: Knopf, 1949. Pp. 459–468.
4. Educational Extension Division, State Geological Survey, Department of Registration and Education, State of Illinois. *Mineral industries of Illinois.* (Map) Springfield: Author, 1955.
5. EMPEY, L. T. Social class and occupa-

tional aspiration: A comparison of absolute and relative measurement. *Amer sociol. Rev.,* 1956, **21,** 703–709.

6. Lancaster, H. O. Complex contingency tables treated by the partition of chi-square. *J. Royal stat. Society,* Series B, 1951, **13,** 242–249.

7. Mollenkopf, W. G., & Melville, S. D. A study of secondary school characteristics as related to test scores. *Research Bulletin No. 56–6.* Princeton: Educational Testing Service, 1956.

8. Thorndike, R. L. Community variables as predictors of intelligence and academic achievement. *J. educ. Psychol.,* 1951, **42,** 321–338.

9. U. S. Bureau of Census. *U. S. census of agriculture:* Vol. 1, Counties and state economic areas, Part 5, 1954. Washington: U. S. Government Printing Office, 1956.

10. Warner, W. L., Havinghurst, R. J., & Loeb, M. B. The social role of the teacher. In T. M. Newcomb, & E. L. Hartley, (Eds.) *Readings in social psychology.* New York: Henry Holt, 1947. Pp. 479–480.

Received October 24, 1957.

THE JOURNAL OF

EDUCATIONAL PSYCHOLOGY

Volume 49 April, 1958 Number 2

DIFFERENTIAL RETENTION OF COURSE OUTCOMES IN EDUCATIONAL PSYCHOLOGY

WILLIAM P. MCDOUGALL

Washington State College

One of the paramount problems of all educational endeavor is that of making the learning experiences of students more lasting. Though the problem of retention has been studied in many different school subjects, relatively little research has been reported dealing directly with the permanency of different kinds of course outcomes. The need for such evidence is suggested by the following quotation from the *Taxonomy of Educational Objectives.*

> For the most part research on the problems in retention, growth and transfer has not been very specific with respect to the particular behavior involved. Thus, we are not usually able to determine from this research whether one kind of behavior is retained for a longer period of time than another or which kinds of educative experiences are most efficient in producing a particular kind of behavior. Many claims have been made for different educational procedures, particularly in relation to permanence of learning; but seldom have these been buttressed by research findings (**2,** p. 23).

It was the purpose of this study to measure retention of different course outcomes in a beginning course in educational psychology. The outcomes examined included: (*a*) knowledge and the intellectual abilities and skills, (*b*) translation, (*c*) interpretation, and (*d*) extrapolation. These objectives were defined by the *Taxonomy of Educational Objectives* (**2**), a handbook consisting of a logical and psychological classification of educational goals. This handbook enables test constructors to define very clearly the classes of behavior being measured in that it provides extensive definitions together with examples of test situations measuring the various behavioral objectives.

PROCEDURE

The general plan of the study involved the construction of tests to measure a variety of educational objectives in a beginning course in educational psychology at the University of Nebraska. The course, Human Behavior and Development, is the second of a two-course sequence taken by teacher trainees. It encompasses primarily the content areas of learning and evaluation. For this study, the content considered was delimited to the materials studied about tests and measurements in order to permit more intensive and uniform sampling of the objectives.

The tests were related to the course by using the course syllabus and accompanying references which were used by all instructors teaching the various sections of the course. For each of the objectives tested, a few examples of items patterned after the "Taxonomy" definitions follow:

Knowledge

1. Which of the following is most easily measured by a test: (*a*) problem-solving ability, (*b*) study skills, (*c*) factual information, (*d*) ability to comprehend.

2. Which of the following is an individual intelligence test: (*a*) California Test of

Mental Maturity, (*b*) Stanford Binet, (*c*) Ohio State Psychological Test, (*d*) Primary Mental Abilities.

3. A test that places minor emphasis on the time limit is called a: (*a*) diagnostic test, (*b*) performance test, (*c*) survey test, (*d*) power test.

4. Which of the following would be of most value in determining the typical behavior of a student: (*a*) observation, (*b*) projective testing, (*c*) individual intelligence testing, (*d*) school achievement records.

Item 1 is designed to measure knowledge of specific fact, Item 2, knowledge of a classification, Item 3, knowledge of terminology, and Item 4, knowledge of methodology.

Translation

1. A major use of testing is for diagnosis. Which of the following test situations represents the best example of the foregoing statement? (*a*) a comprehensive achievement battery at the end of high school, (*b*) an achievement battery given early in the year, (*c*) an intelligence test, (*d*) a series of tests used to determine a student's grade.

2. If Bill scored at the 88th percentile in Social Service on the Kuder Preference Test, it would indicate that: (*a*) Bill got 88% of the answers correct, (*b*) he has more ability in Social Service than 88% of his norm group, (*c*) only 12% of the norm group showed more interest in Social Service than he did, (*d*) that 88 out of 100 will do better than he did on this test.

The first exercise involves translation of a formal statement by requiring the student to identify a concrete example. The second item involves the translation of quantitative data to its corresponding verbal meaning.

Interpretation

Data are given below on five pupils enrolled in a class of 30 ninth graders. The test data are based on performance at the end of the first semester. Read over the summary and then show which pupil each statement best fits by marking the pupil's number on the answer sheet.

Pupil	IQ	Arith.	Calif. Ach. Test Performance Read.	Lang.	Teacher's Estimate of Ach. Rank in Class
1	88	9.1	8.0	8.3	20
2	99	9.7	9.6	9.5	14
3	132	9.5	9.8	10.2	12
4	138	11.8	12.3	12.0	3
5	101	10.0	10.1	10.9	4

1. The pupil who should be doing considerably better in his school achievement.

2. The accuracy of the IQ seems most doubtful in which case?

3. A bright student making good use of his ability.

4. Teacher regards abilities too highly according to test results.

5. Teacher's rank most consistent with test scores.

Each of the foregoing situations involves the ability to deal with a configuration of ideas or data recognizing the relationship and relative importance of each. The inferences or generalizations made from the data do not extend beyond the data but are confined to the material presented.

Extrapolation

The five students for whom the data are given below are in kindergarten. These test data are based on test performance at the beginning of the second semester. After examining the data, indicate which pupil best fits each of the following statements by marking the number of the student on the answer sheet.

Student	CA	MA on Stanford Binet	Percentile Rank on Readiness Test
1	5–10	7–4	72
2	6–4	5–4	22
3	5–10	5–5	64
4	5–8	5–6	45
5	5–6	6–10	38

Which student:

1. Is apparently in need of stimulating experiences but has fairly high aptitude?

2. Apparently comes from a very stimulating environment?

3. Is most characteristic of the average for this group?

4. Can you predict will have the lowest ability three years from this time?

The first two situations require the student to extend the implications of the data to another topic or situation. The third situation requires extension from a sample to a universe. The last item involves time dimension and requires prediction on the basis of the data presented.

The tests were then administered on a trial basis to a group of 75 educational psychology students who had completed units on tests and measurements. The tests were then analyzed, refined, and used as instruments to study the retention of the different course outcomes. The refined tests were given as a pretest, a test at the completion of the course, and a retest approximately four months later. There were 301 students who took both the pretest and the test, and 172 of this group took the retest. This latter group was used in the study of retention.

The appropriateness of the tests involved in the study was examined after the test was administered to the trial test group and again after the test had been revised.

The original trial tests contained approximately 30 items measuring each objective. The curricular validity of items was established by agreement among several instructors teaching the course involved in the study. Pooled judgment of several instructors was also used to assure that each item was correctly matched with the corresponding "Taxonomy" definition. The items were then studied after administration to the trial test group. Item difficulty and item discrimination were determined and substandard items were dropped or revised. Evidence of ambiguity in items and ineffective distractors were also studied and many items were revised or eliminated on this basis. The resulting refined tests contained approximately 24 items each, and when combined required approximately one and one-half to two hours for administration.

The tests were studied again at the time of the second testing in the retention experiment. At this time, 310 people took the test as part of their course final examination. Item difficulty, item discrimination, and test reliability were determined. Homogeneity of behavior measured by the different tests was studied in two ways: First, the correlations of the items with their respective test totals were compared with item correlations using the total of the four tests combined as a criterion. Second, an F test for departure from homogenity proposed by Neidt (**10,** p. 390) was applied. This latter technique indicated whether or not there is a relatively greater lack of homogeneity between or among areas than within areas measured. The semiexternal criterion of course marks was correlated with the scores to further establish test validity. The correlation coefficients between each test and a measure of scholastic aptitude, the L score on the American Council on Education Psychological Examination, was computed to determine the degree to which verbal ability was present in each of these tests.

In the study of retention, the suitability of the sample of 172 students who took the retest was determined by comparing the performance of this group with the performance of the group who did not take the retest. The degree of relationship between the scores on each test administration was found by computing correlation coefficients between the pretest and the test, the test and retest, and the pretest and retest. The differences between the means of scores on each of the test administrations was determined and tested for significance. Retention was then studied by computing the average percentage of gain retained for each of the separate objectives measured.

RESULTS

Analysis of the Tests

The test item difficulty, reported in terms of percentage of the group who responded correctly to the item, was determined for all tests. The mean level of difficulty for the knowledge test was 62.13%. The means for translation, interpretation, and extrapolation were 60.45, 60.61, and 56.52, respectively. The individual difficulty percentages tended to cluster about the means and seemed to be well distributed with no items either being answered correctly or missed by 100% of the group.

Item discrimination was determined by correlating each item with the total test score. To obtain these correlations, the upper and lower 27% of the distribution are designated as the criterion variable, and by entering the appropriate percentages in an item analysis table (**4**), the correlations may be estimated. Such correlations indicate the tendency for students who make high scores on the total test to mark the individual item correctly.

On the combined tests, 54% of the total items were found to yield correlations of .40 or above. Twenty-nine per cent of the total items were between .20 and .30. Only fifteen, or 17% of the total items, yielded correlations of less than .20. Two items were found to yield negative correlations and were eliminated from use in the retention study. The above percentages were fairly characteristic of all of the tests with slightly more low-correlation items in the knowledge and translation tests than in the interpretation and extrapolation tests.

The Spearman-Brown and the Kuder-Richardson estimates of reliability are shown in Table 1.

Apparently the small number of items included in each test is the major reason for the somewhat low reliabilities. For evaluating the level of group accomplishment, such reliabilities may be regarded as acceptable according to some sources (**8,** p. 609). Certainly higher reliabilities would be more desirable, but in this experiment the limiting factor of testing time would have made it extremely difficut to include more items in the tests.

TABLE 1

SPEARMAN-BROWN AND KUDER-RICHARDSON ESTIMATES OF RELIABILITY

Test	Number of items	Odd-even correlation	Spearman-Brown Estimate	Kuder-Richardson Estimate
Knowledge	24	.297	.458	.495
Translation	22	.290	.450	.507
Interpretation	23	.477	.646	.531
Extrapolation	24	.440	.611	.537

One positive indication of homogeneity of the behavior measured by the different tests can be obtained by comparing the individual item correlations when using the respective test scores as a criterion with those obtained by using the total scores of the four tests combined as a criterion. Since the total score constitutes the criterion with which the item is compared, the higher the correlation the more the behavior measured by each item is like the behavior measured by the total test. It was noted that when all of the tests were combined into one single test score and the items correlated with this total, most of the correlations were reduced. This reduction would indicate a greater heterogeneity of test content when tests were combined or, conversely, a greater homogeneity of content in the separate tests. It was not possible by this method, however, to determine the degree to which each test is homogeneous with respect to each other test. To test this hypothesis, an F test for departure from homogeneity was applied. Intra- and interarea correlations were obtained and averaged according to the function $\frac{1}{2} \log_e (1 + r)/(1 - r)$ as necessary for substitution into the formula for computing the F values which is:

$$F = \frac{1 + \bar{r}_w - 2\bar{r}_a}{1 - \bar{r}_w}$$

where $\bar{r}_w$ is the average intra-area coefficient and $\bar{r}_a$ is the average interarea coefficient of correlation. The resulting F values are shown in Table 2. Inspection of Table 2 shows that the resulting F values are significant beyond the 1% level of confidence between knowledge and translation, knowledge and interpretation, and translation and extrapolation. The F value for translation and interpretation is significant at the 5% level. The F values between knowledge and extrapolation and interpretation and extrapolation are not significant, although the first of these approaches significance. The hypothesis that behaviors measured by the different tests were homogeneous with respect to each other can be rejected between all tests except knowledge and extrapolation and interpretation and extrapolation.

On the basis of these results the interpretation and extrapolation tests were combined since they did not seem to be performing separate functions. The resulting F values with these two tests combined are shown in Table 3.

Inspection of Table 3 reveals that all values of F are significant beyond the 1% level of confidence. The hypothesis that these three tests measure behaviors homogeneous with respect to each other is rejected. The remainder of the experiment considered interpretation and extrapolation as a single test. The resulting Spearman-Brown estimate of reliability for this test would become .773.

A semiexternal criterion, namely final course marks, was employed to obtain a measure of empirical validity. The resulting correlations between the tests and final grades centered about .60, demonstrating a high positive relationship using such a criterion. These correlations would be spurious to the extent that the tests used in the experiment consituted as much as one-sixth of the final grade.

The correlations between the L score of the American Council on Education Psychological Examination and each test are as follows:

Test	r
Knowledge	.364
Translation	.362
Interpretation-Extrapolation	.343

It is evident from the inspection of these coefficients that the influence of the scholastic aptitude factor as measured by the L score is equally present in the performance required by the different tests.

TABLE 2

VALUES OF F FOR TESTS OF HOMOGENEITY BETWEEN TESTS

Test	Translation	Interpretation	Extrapolation
Knowledge	1.388	1.332	1.178
Translation		1.310	1.423
Interpretation			1.005

Note.—Required for significance, 309 and 309 degrees of freedom, 1% = 1.33
5% = 1.22

TABLE 3

VALUES OF F FOR TESTS OF HOMOGENEITY BETWEEN TESTS

(INTERPRETATION-EXTRAPOLATION COMBINED)

Test	Translation	Interpretation-Extrapolation,
Knowledge	1.388	1.358
Translation		1.489

Note.—Required for significance, 309 and 309 degrees of freedom, 1% = 1.33
5% = 1.22

The Study of Retention

The differences on scores of the 172 students who took the retest and those who did not were determined and tests of significance applied. It was established that this sample was characteristic of the population of 301 from which it was taken.

The possibility that subject matter learned in other courses might transfer was also considered. It was discovered that none of the students who participated took courses during the retention period that dealt systematically with the area of tests and measurements. Apparently it was safe to conclude that only incidental amounts of transfer, if any, would be expected.

The relationship between the three administrations of each test was studied. The resulting correlation coefficients were positive in all cases but not to a high degree. The values ranged from .274 to .599 and averaged about .44. Such correlations indicated that individuals tended to maintain their relative rank on the successive test administrations. The correlations were slightly higher for the interpretation-extrapolation test than for the others, with an average correlation of .542.

To determine if the differences in mean performance on the various test administrations were significant, a *t* test for correlated data was applied. In Table 4 the differences, together with the accompanying *t* values, are shown.

It may be noted from inspection of Table 4 that all of these differences are significant at the 1% level except the difference between the pretest and retest for the interpretation-extrapolation test. This difference is significant at the 2% level. These results show that, on the average, a significant amount of material was learned during the instruction period, a significant amount forgotten during the four-month retention period and at the end of the retention period the students still retained enough learning so that their performance was significantly different from that at the time of the pretest.

The amount of material retained for each test may also be reported in terms of percentage of gain retained. These percentages are also reported in Table 4.

To determine if the differences between the percentages were significant, a *t* test for correlated data was applied. The difference of .78% between knowledge and translation yielded a *t* value of .222 which is not significant. The percentage difference between knowledge and interpretation-extrapolation was 6.55 with an accompanying *t* of 1.985 which is significant at the 5% level. The difference of 5.77%

TABLE 4

Mean Scores, Differences Between Mean Scores with Accompanying *t* Values, and Percentage of Gain Retained for the Three Administrations of the Tests

	Test		
	Knowledge ($N = 172$)	Translation ($N = 172$)	Interpretation-Extrapolation ($N = 172$)
Pretest Mean (M_P)	11.90	11.05	23.13
Test Mean (M_T)	15.55	13.83	28.31
Retest Mean (M_R)	14.55	13.09	27.23
$M_T - M_P$ (Gain)	3.65	2.78	5.18
t	10.42	13.29	12.33
$M_T - M_R$ (Loss)	1.00	.74	1.08
t	3.33	2.74	2.57
$M_R - M_P$ (Gain Retained)	2.65	2.04	4.10
t	10.19	8.16	9.11
$\frac{M_R - M_P}{M_T - M_P}$ (Percent of Gain Retained)	% = 72.60	% = 73.38	% = 79.15

Note.—Required for Significance, 171 Degrees of Freedom, 1% = 2.58
2% = 2.32

between translation and interpretation-extrapolation yielded a *t* of 1.748 which is significant at the 10% level of confidence.

The course of learning and retention for each behavior studied may also be expressed in terms of percentage of items answered correctly at each testing.

The greatest gain and relatively the greatest loss were made on the knowledge test, the percentage of items correct increasing during the course from 49.6% to 64.8% and dropping off to 60.6%. The average percentage of items correct for translation began with 50.2%, increased to 62.4% and dropped to 59.5%. The corresponding percentages for interpretation-extrapolation were 49.2%, 60.2%, and 57.9%.

Discussion

The results of this study indicate the need for carefully delineated course objectives. The homogeneity analysis in this experiment showed that tests constructed to measure certain behavioral outcomes apparently perform separate functions as evaluation devices. Thus, to insure that multiple course outcomes, in line with the objectives of instruction, are achieved, it becomes necessary to design evaluation instruments to accomplish these separate functions. These results tend to agree with the results of previous studies done by Tyler (**13**), McConnell (**9**), Johnson (**7**), Brown (**3**), Horrocks (**6**), and Bedell (**1**), all of which make it apparent that the achievement of one objective cannot be inferred from the achievement of another. Remmers has expressed this point when he concluded (**11**, p. 31): "... the educator must clearly define each objective in terms of the measure of its attainment. The attainment of a particular objective cannot be inferred from measured attainment of another objective."

The majority of studies reported in the literature suggest much of what is learned in school is forgotten. It has long been the concern of educators to provide learning experiences of more permanent value. From this standpoint, the results of this investigation suggest that increased emphasis on some of the higher levels of understanding such as interpretation and extrapolation will lead to more economical learning. As the authors have defined the objectives in the "Taxonomy," each higher level of intellectual ability is built on and includes the previous levels. To emphasize such abilities as interpretation and extrapolation means that the possession of knowledge and the ability to translate it will be a part of the learning experience, but that understanding will go beyond these lower levels of intellectual endeavor and involve mastering more permanent abilities and skills. Such practices have not always been the case. Tyler (**13**) found that interviews with college students indicated that more than 60% of the students in college believe their chief duty is to memorize information. Tyler stated that the emphasis given to recall of fact in the typical college examination is one of the chief reasons for the existence of this belief.

It has been previously shown in studies done by Tyler (**12**), Wert (**14**), and Frutchey (**5**) that such outcomes as the ability to apply principles to new situations and interpret new experiments demonstrated much higher degrees of permanency than abilities involving only the recall of specifics. The results of the present experiment agree in general with what has been previously done. It also suggests that such a device as the "Taxonomy" will enable us to do a far more systematic and communicable job in studying different outcomes of instruction.

Summary

The purpose of this experiment was to study the differential retention of certain course outcomes in a beginning educational psychology course.

Tests were constructed to measure four different behavioral outcomes in the content area of tests and measurements. These outcomes were: (*a*) knowledge, (*b*) translation, (*c*) interpretation, and (*d*) extrapolation. As a result of a homogeneity analysis of the behaviors measured by these different tests, it was found desirable to combine the interpretation and extrapolation tests in that they seem to be performing a similar measurement function.

The tests were administered as a pretest before the units on tests and measurements were studied, at the completion of the units, and a third time after approximately four months had elapsed. The results of the study of retention indicated that the abilities to interpret and extrapolate were retained to a significantly greater degree than the ability to recall knowledge or translate this knowledge from one form to another. It was concluded that there was differential retention among the behavioral objectives measured.

REFERENCES

1. Bedell, R. C., The relationship between the ability to recall and the ability to infer in specific learning situations, *Bull. of the Northeast Missouri State Teachers Coll.*, XXXIV, No. 9, Kirksville: Northeast Missouri State Teachers College, 1934.
2. Bloom, B. S., (Ed.) *Taxonomy of educational objectives,* New York: Longmans, Green, 1956.
3. Brown, Clara M., *Evaluation and investigation in home economics,* New York: F. S. Crofts, 1941.
4. Fan, Chung-Teh, Item Analysis Table, Princeton, New Jersey: Educational Testing Service, 1952.
5. Frutchey, F. P., Retention in high school chemistry, *J. higher Educ.*, 1937, **8,** 217–218.
6. Horrocks, J. E., The relationship between knowledge of human development and the ability to use such knowledge, *J. appl. Psychol.*, 1946, **30,** 501–508.
7. Johnson, P. O., Differential functions of examinations, *Studies in College Examinations,* Minneapolis: Univer. Minnesota, 1934, pp. 43–50.
8. Lindquist, E. F. (Ed.), *Educational measurement,* Washington, D. C.: American Council on Education, 1951.
9. McConnell, T. R., A study of the extent of measurement of differential objectives of instruction, *J. educ. Res.*, 1940, **33,** 662–670.
10. Neidt, C. O., Technique for testing the homogeneity of separately-evaluated behavior characteristics, *Iowa State J. Sci.,* Fall, 1949, 390–392.
11. Remmers, H. H., & Gage, N. L., *Educational measurement and evaluation,* New York: Harper, 1955.
12. Tyler, R. W., Permanence of Learning, *J. higher Educ.*, 1933, **4**, 203–204.
13. Tyler, R. W., The relation between recall and higher mental processes. In C. H. Judd, *Education as cultivation of the higher mental processes,* New York: Macmillan, 1936.
14. Wert, J. E., Twin examination assumptions, *J. higher Educ.*, 1937, **8,** 136–140.

Received October 7, 1957.

Journal of Educational Psychology
Vol. 49, No. 2, 1958

ATTITUDES TOWARD SCHOOL OF HIGH SCHOOL PUPILS FROM THREE INCOME LEVELS

JOHN K. COSTER

Department of Education, Purdue University

Educators have become increasingly concerned with the relationship between social status and educational outcomes. Numerous studies of this relationship have been reported. The findings have demonstrated that social status is related to practically all educational experiences.

The relationship between social status and intelligence and achievement has been emphasized in these studies (**7, 14**). Other studies have involved personality (**2, 6**), extracurricular activities (**9, 15**), social acceptance (**1, 11**), honors received (**1**), attitudes (**8, 10**), and morale (**3**). Results usually indicate that upper status pupils exceed lower status pupils on achievement and intelligence test scores, marks in school, adequacy of adjustment, number of activities, social relationships, and attitudes toward school. The differences are generally statistically significant.

In the present study, the relationship between specific attitudes toward school and level of income was investigated. The purpose was to ascertain on which of a number of attitudinal items pupils varied in their responses when they were divided into three income groups.

Procedure

A questionnaire, containing a morale scale and a "house and home" scale, was administered to approximately 3,000 pupils in nine central and south central Indiana high schools. The morale scale, constructed as part of another study (**3**), contained 27 attitudinal items. The items pertained to the school, teachers, school program, other pupils, and the value of education. Each item in the scale was stated as a question, and was followed by a list of five possible responses. The responses reflected (*a*) a very favorable attitude, (*b*) a favorable attitude, (*c*) a neutral (neither favorable nor unfavorable) attitude, (*d*) an unfavorable attitude, and (*e*) a very unfavorable attitude. Following is an example of a typical item and list of responses.

Item: What is your general opinion of the other boys and girls in your high school?

——a. They are the *best* group of boys and girls in the world!

——b. I feel that we have a *good* group of boys and girls in our high school.

——c. Some of the other students are all right; some are not.

——d. I feel that this high school has a *poor* group of boys and girls.

——e. They are the *worst* group of boys and girls in the world!

Pupils were instructed to check the responses with which they agreed most closely.

An indication of income level was obtained from a "house and home" scale. This scale listed seven things either found in the home or provided for the pupil. The items were a vacuum cleaner; an electric or gas refrigerator; a bath tub or shower with running water; two automobiles (excluding trucks); lessons in drama, art, expression, dancing, or music provided outside of school; an automatic dishwasher; and a cabin or cottage for vacations. Pupils checked the items which applied to them.

The "house and home" scale has been used extensively by Remmers and others (**12**) in the *Purdue Opinion Panel* studies to divide pupils into income groups (e.g., **10**). Elias (**5**) and Remmers and Kirk

TABLE 1

NUMBER AND PERCENTAGE OF PUPILS WHO CHECKED ITEMS ON HOUSE AND HOME SCALE, AND NUMBER AND PERCENTAGE OF PUPILS IN EACH INCOME GROUP

Number of items checked	Number and percentage of pupils checking				Income Group
	N	%	*N*	%	
0	11	1.2			
1	63	7.2	219	24.9	Low
2	145	16.5			
3	270	30.8			
4	288	32.8	558	63.6	Middle
5	83	9.5			
6	9	1.0	101	11.5	High
7	9	1.0			
Total	878	100.0	878	100.0	

(13) have reported on the validity of the scale.

A sample of 878 cases was selected from the returns. The sample included 100 questionnaires, selected randomly, from the six larger schools, and all useable questionnaires from the three smaller schools.

Based on the number of items checked on the "house and home" scale, *S*s were divided into three income groups. The high income group included *S*s who checked five to seven items. The middle income group included those who checked three or four items. And the low income group included those who checked two or fewer items. The number and percentage of pupils who checked each number category, and the number and percentage of pupils in each income group are shown in Table 1.

Responses to each attitudinal item were tabulated by income group. The responses were then combined into two categories. One group included favorable and very favorable responses. The second group included all other responses.

For each item, the following null hypothesis was postulated: There is no difference in the responses of pupils of varied income groups. Each of the 27 hypotheses was tested by the chi-square technique. The tests were based on a series of 2 × 3 contingency tables. The combination of responses provided a uniform series of tables, with a minimum expected frequency of five in each cell.

RESULTS AND DISCUSSION

The results of the chi-square tests are given in Table 2. The table also shows the percentage of pupils, by income group, who checked favorable and very favorable responses. The item-questions were abridged to conserve space in the table. The column headed "P" indicates the probability level associated with chi-square values.

The items were divided into seven groups to facilitate interpretation: (A) Attitudes Toward Teachers, (B) Attitudes Toward the School, (C) Attitudes Toward School Program, (D) Attitudes Toward Appropriateness of School Work, (E) Attitudes Related to Future Expectations, (F) Attitudes Related to Social Acceptance, and (G) Miscellaneous Attitudes. The letters are used in designating the items in Table 2.

The data show that responses varied significantly on relatively few items. Only eight of the 27 hypotheses could be rejected, six at the 1% level and two at the 5% level. The responses among groups differed widely, ranging from practically no variation to extremely significant variations. The frequencies generally varied

TABLE 2

RESULTS OF TESTS OF SIGNIFICANCE SHOWING PERCENTAGE OF PUPILS CHECKING VERY FAVORABLE AND FAVORABLE RESPONSES, BY INCOME GROUPS

Item		Income Group			Total	*P*
		High	Middle	Low		
A-1	What is your opinion of your high school teachers?	69.3	61.6	65.8	63.6	.30
A-2	Do your teachers treat you fairly?	89.1	84.8	83.1	84.9	.50
A-3	Are your teachers personally interested in you?	64.4	55.4	49.3	54.9	.05
A-4	Do your teachers "know" and understand their subjects?	82.2	83.7	83.6	83.5	.95
A-5	How well are your subjects taught?	54.5	52.5	50.7	52.3	.90
A-6	Do your teachers help you sufficiently with your school work?	84.2	78.9	74.0	78.2	.20
A-7	Would you ask adults in your school for help with personal problems?	37.6	31.2	35.2	32.9	.50
B-1	What is your general opinion of your high school?	82.2	74.0	65.8	72.9	.01
B-2	How well is your school organized?	67.3	69.2	68.9	68.9	.95
B-3	How satisfactory are the working and studying conditions?	38.6	40.7	38.8	40.0	.90
B-4	How satisfactory are the equipment and facilities?	34.7	29.4	26.5	29.3	.50
B-5	How satisfactory is the grading system?	80.2	79.6	81.7	80.2	.80
B-6	What is your opinion of the school spirit in your school?	53.5	51.8	49.8	51.4	.90
C-1	What is your opinion of the group of subjects your school offers?	65.3	62.9	54.8	61.2	.10
C-2	What is your opinion of the number of activities in your school?	48.5	52.5	50.2	51.5	.70
D-1	Is your school work the kind of work you like to do?	71.3	59.5	61.6	61.4	.10
D-2	Is your school work interesting?	86.1	75.1	73.5	76.0	.05
E-1	Will your school work be useful after you leave school?	94.1	91.2	88.1	90.8	.20
E-2	Will going to high school help you get more satisfaction from living?	94.1	90.1	85.8	89.5	.10
E-3	What are your chances of getting the job you want after high school?	70.3	59.9	45.2	57.4	.001
F-1	Are you satisfied with your social life in high school?	74.3	71.0	60.3	68.7	.01
F-2	Do the other students like you?	85.1	80.5	67.1	77.7	.001
F-3	How well do other people in your school treat you?	83.2	83.7	76.7	81.9	.10
F-4	What is your opinion of the other boys and girls in your school?	71.3	65.2	49.8	62.1	.001
G-1	Are your parents interested in your high school work?	96.0	92.1	76.7	88.7	.001
G-2	How do people in your community feel about your high school?	73.3	73.5	66.2	71.6	.20
G-3	How hard are you working or studying in high school?	46.5	48.9	46.1	48.0	.80

greatest on items which involved interpersonal relationships. And they appeared to vary least on items which the pupils could consider objectively, with limited emotional attachment.

Significant variations were noted for three of the four items in the social acceptance group (Group F). Low income pupils reacted less favorably than other pupils to their social life (Item F-1), to being liked by other pupils (F-2), and to other pupils (F-4). According to unpublished data (**4**), high and middle income pupils significantly exceeded others in the percentage who associate with fellow pupils outside of school. Low income pupils were more likely to associate with youth from other schools or not in school.

The attitudes on social acceptance seemed to be related to other attitudes on which differences in responses were observed. Low income pupils apparently are not as sure of parental interest in school work as other pupils (G-1). Whereas practically all high income pupils indicated that they felt that their parents were interested in their work, only three fourths of the low income pupils expressed similar opinions. These differences were highly significant, with $P < .001$.

Low income pupils also differed from other pupils in their estimates of the personal interest of their teachers (A-3). The differences were significant at the 5% level. This item was the only one of the seven items pertaining to teachers on which responses varied significantly.

The item on general impression of the high school (B-1) was the only item related to teachers, school, school program, appropriateness of work, and value of education on which differences were significant at the 1% level. In view of the homogeneity of other items, it would seem that responses to this item were affected more by the nature of relationships than by the nature of the school and school program.

Responses varied widely to the item on future employment (E-3). Over two thirds of the high income pupils, as compared with less than one half of the low income group, expressed favorable responses about getting the kind of job they want. Differences were significant at the 0.1% level. This item may be related to post high school educational aspirations. It was found that one half of the high income pupils in the sample plan to go to college, as compared with less than one sixth of the low income group (**4**).

Except for the two items mentioned previously, responses to items on teachers (Group A) and school (Group B) varied slightly or not at all. Pupils in the three income groups were virtually in complete agreement on items related to the technical operation of the school and the technical competency of teachers.

The responses to items on school program (Group C), appropriateness of school work (Group D), and the value of education (E-1 and E-2) generally varied more than the responses to items on teachers and the school. The responses of high income pupils were more favorable for five of the six items in these groups, but, except for item D-2, variations were not significant. High income pupils were more interested in their school work than others (D-2), and differences were significant at the 5% level. The *S*s responded uniformly to the number of activities in the school (C-2), even though low income pupils participated in significantly fewer activities (**4**). The pupils in all groups reacted favorably to the value of education. The low income pupils, however, reacted more favorably to the utility value of education (E-1) than to the enrichment value (E-2).

The low income pupils differed from others—but not significantly—on estimates of how people in their communities felt about their high schools (G-2). And on the question of how hard pupils were working in high school (G-3), no variation among groups was observed.

Conclusions

The data seem to support the following conclusions:

1. Responses of pupils of different income levels were more likely to vary on items related to interpersonal relationships than on items which involved an objective appraisal of the school or the school program.

2. The schools in the study have provided an educational program uniformly accepted by pupils of the three income levels. They have been less successful in integrating all pupils into the social structure of the school. How acceptance may be gained for all pupils is undoubtedly a perennial problem.

3. The low income pupil is less likely to enjoy strong parental interest and support than other pupils. An immediate, practical problem confronting the schools, therefore, is stimulating interest of all parents in school and school work.

4. Variations in estimates of possible satisfactory future employment among pupils of varied income levels suggest that more attention should be given to helping noncollege, low income pupils select, prepare for, and enter an appropriate vocation.

Summary

When 878 pupils from nine Indiana high schools were divided into three income groups, it was found that they responded similarly to attitudinal items on school, school personnel, school program, and the value of an education. The responses varied significantly with income level, however, on items related to interpersonal relationships. The items on which differences were observed pertained to social life, being liked by other pupils, opinions of other pupils, feelings of parental interest in school work, and personal interest of teachers. Although pupils responded uniformly on specific items pertaining to the school, they varied significantly, according to income level, in their general impression of their schools. They also varied significantly in their estimates of being able to get the kind of jobs they want after they leave school.

References

1. Abrahamson, A., Our status system and scholastic rewards. *J. educ. Sociol.*, 1952, **25,** 441–50.
2. Coleman, H. A., The relationships of socio-economic status to performance of junior high school students, *J. exp. Educ.*, 1940, **9,** 61–63.
3. Coster, J. K., Factors related to morale in secondary schools. Unpublished doctoral dissertation, Yale Univer., 1955.
4. Coster, J. K., Characteristics of pupils of three income levels. Lafayette, Indiana; Purdue University, 1958. Unpublished manuscript.
5. Elias, G., A study of certain methods of attitude measurement and related variables. Unpublished master's thesis, Purdue Univer., 1944.
6. Gough, H. G., Relationship of socio-economic status to personality inventory and achievement test scores, *J. educ. Psychol.*, 1946, **37,** 527–540.
7. Havighurst, R. J., & Breese, H. F. Relation between ability and social status in a midwestern community. III. Primary mental abilities, *J. educ. Psychol.*, 1947, **38,** 241–247.
8. Hieronymus, A. N., Study of social class motivation: Relationships between anxiety for education and certain socio-economic and intellectual variables, *J. educ. Psychol.*, 1951, **42,** 193–205.
9. Hollingshead, A. B., *Elmtown's youth.* New York: John Wiley, 1949.
10. Kirk, R. B., Attitudes toward public education as related to N variables. Unpublished doctoral dissertation, Purdue University, 1952.
11. Morgan, H. G., Social relationships of children in a war-boom community, *J. educ. Res.*, 1946, **40,** 271–86.

12. Remmers, H. H., (Ed.), *Purdue opinion panel.* Lafayette, Indiana: Division of Educational Reference, Purdue Univer.

13. Remmers, H. H., & Kirk, R. B., Scalability and validity of the socio-economic status items of the Purdue Opinion Panel, *J. appl. Psychol.*, 1953, **37,** 384–386.

14. Shaw, C., The relation of socio-economic status to educational achievement in grades four to eight, *J. educ. Res.*, 1943, **37,** 197–201.

15. Smith, H. P., A study of the selective character of American education: Participation in school activities as conditioned by socio-economic status and other factors, *J. educ. Psychol.*, 1945, **36,** 229–246.

Received July 28, 1957.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 2, 1958

VALIDATION OF NEW ITEM TYPES AGAINST FOUR-YEAR ACADEMIC CRITERIA

JOHN W. FRENCH

Educational Testing Service

The College Entrance Examination Board undertook in 1951 a study to explore the effectiveness of a series of new aptitude tests that might prove to be contributive supplements to the Scholastic Aptitude Test or effective substitutes for parts of it.

Validity studies of the SAT itself are carried out routinely. Substantially all of these use as criteria the grades received during freshman year. This has been done mainly because of the great delay encountered in waiting for the longer-term criteria. Furthermore, while students take a considerable variety of courses in freshman year, their freshman programs are much more alike than their upperclass programs. For this reason "average freshman grades" may be not only more quickly available but also more meaningful than average grades received when the students are working in different subject-matter areas having different degrees of difficulty.

It is useful to consider some hypotheses for the change that might occur in the validity of an aptitude test between freshman and upperclass years in college. The following conditions should lead to a *decrease* in the validity of aptitude tests:

1. Seniors take more varied courses than freshmen; success in the various courses, some easy and some difficult, will be hard to predict.
2. Time between testing and the measurement of the criterion allows more scope for changes to take place in the individual students as a result of different experiences or different rates of maturation.
3. Attrition at college cuts down the range of ability between freshman and senior years.

Conditions possibly leading to an *increase* in validities are as follows:

1. A lack of adequate adjustment to college life in freshman year might introduce extraneous influences on scholastic success.
2. More uniformly high motivation and a more serious attitude toward work in upperclass years may cut down one source of extraneous variance.
3. Emphasis on memory work in freshman year may depend on motivation or other factors, while the understanding and problem solving required in upperclass years may depend more upon the aptitudes measured by most test scores.

It is not easy to guess at the resultant of such factors as these.

There have been very few studies where validities of High School Record and of College Board and other tests for freshman grades have been compared with validities for four-year grades. Studies by Dwyer (**2**), Brush (**1**), and Frederiksen (**3**) have shown, in general, that four-year cumulative average validities do not differ consistently from freshman validities. Findings in the present study generally confirm this conclusion.

Even less has been done in validating the SAT against major field grades. An unpublished study carried out at Stanford University (**6**) shows validities of the SAT and high school record for cumulative average and major-field grades. The superiority of the high school record in that study and the sex differences found for the validity of the SAT for Social Science grades are not confirmed by findings in the present study.

THE EXPERIMENTAL TESTS

In addition to the High School Record, SAT-V (verbal), SAT-M (mathematical), and the CEEB English Composition Test, the measures investigated in this study consisted of 11 newly adapted or newly developed aptitude tests, some with part scores. Descriptions of the tests and reliabilities by the Kuder-Richardson formula No. 20 are as follows:

1. *Social Studies Reading.* A 1,000-word passage by Hamilton concerning the Bill of Rights with questions on interpretation, vocabulary in context, and the structure of the passage. 25 four-choice items, 25 minutes. Reliability, .66.

2. *Science Reading.* A 1,200-word essay on "A Piece of Chalk" by Huxley with questions on interpretation, vocabulary in context, and the structure of the passage. 25 four-choice items, 25 minutes. Reliability, .71.

3. *Inductive Reasoning.* A spiral omnibus test using items drawn from verbal, nonverbal, arithmetic, science, and social studies materials. The items were of three types found to measure inductive reasoning: analogies, series, and categories or "belonging" items. The great variety of item types was introduced so that the subjects could not develop a uniform approach or uniform method of solution, which would render the test deductive instead of inductive. 65 items, 25 minutes. Reliability, .82.

4. *Integration.* This test was similar to conventional "artificial language" tests except that the rules for translation were more complex, and there was no premium on quick memory. It was developed as a test of one of the factors called "integration" in the Army Air Force Aviation Psychology Program (4). This is the ability to understand and follow complex directions. 15 items, 25 minutes. Reliability, .73.

5. *Sufficiency of Data.* Each problem consisted of a question followed by two mathematical or quantitative facts. The task was to decide whether either fact, both together, both separately, or none were sufficient to answer the question. 30 problems, 25 minutes. Reliability, .80.

6. *Data Interpretation.* This test consisted of statements related to the content of two sets of data: a table on the expenditures of state governments and a verbal exposition of a research concerning enlargement of the thyroid gland. The task was to decide whether the data were sufficient to make each statement true, probably true, false, probably false, or none of these. 30 items, 25 minutes. Reliability, .68.

7. *Visualization.* Drawings indicated how a square sheet of paper was folded and then punched one or two times. The task was to select from five drawings the one that showed how the paper would look when opened. 20 items, 25 minutes. Reliability, .85.

8. *Best Arguments.* Situations involving some sort of dispute were described in a paragraph. Subjects select one or two statements constituting the best argument for each side. Four situations totalling 21 items, 25 minutes. Reliability, very low. (K.R. 20 was not applicable, because the items were not independent from each other.)

9. *Perceptual Speed and Carefulness.* The two parts of this test each contributed to the measurement of Perceptual Speed and Carefulness. (*a*) *Cancellation.* A page of random capital letters typed close, lines single spaced, and reproduced in red. The task was to draw an X over every A. Three minutes were allowed. (*b*) *Picture Discrimination.* Each item consisted of three simple drawings of a face, two exactly alike, and one different in some respect. Three minutes were allowed. The score for *Perceptual Speed* was the number of A's cancelled plus the number of faces correctly marked. Reliability, .94. The *Carefulness* score was the inverse of a score developed by adding omissions on *Cancellation* to five times the wrongs on *Picture Discrimination.* (This scoring formula operated to weight the two parts equally in the total score.) Reliability, .55.

10. *Memory.* This test had 3 parts scored as separate variables for validation purposes: (*a*) *Picture Memory.* A picture of a Venetian palace was studied for five minutes. Later a second picture was presented showing the same palace with some features changed. The students were allowed five minutes to answer 30 true-false questions comparing the pictures. (*b*) *Verbal Memory.* A one-page description of the peoples of Honduras was studied for five minutes. Later the students were allowed five minutes to answer 30 true-false questions about the passage. (*c*) *Number Memory.* Some prices and inventory numbers in department stores were studied for five minutes. Later the students were allowed five minutes to answer 15 five-choice questions calling for recognition of the memorized numbers. The memory portion of each of these parts took place during the initial 15 minutes of the two-hour testing session, and the response portions took place during the last 15 minutes with one and a half hours of other testing

coming in between. Reliabilities for the three parts were respectively .74, .54, and .69.

11. *General Information.* This test presented five-choice factual information items drawn from various fields with the intention of measuring interest in those fields. The items were selected so as to avoid information that would be acquired in school, but to include information that would be gained through hobby work or incidental reading, presumably of the student's own choosing. The scores (number of items answered correctly out of the 15 for each of seven fields) were treated as separate variables in the validation study. The fields included were: (*a*) *Art and Architecture*, (*b*) *Literature*, (*c*) *Social Work*, (*d*) *Government*, (*e*) *Biological Science*, (*f*) *Physical Science*, and (*g*) *Mechanical*. Total items 105; total time 40 minutes. Reliabilities for the parts were respectively .53, .52, .30, .52, .52, .56, and .52.

Administration of the Tests

Ten liberal arts colleges, all of which require the SAT for entrance, participated in the study by scheduling two hours of testing for their entering freshmen and by supplying all course grades and sometimes the high school record. Four tests were administered at each of the colleges by combinations taken so as to provide a substantial number of cases for the most interesting of the intercorrelations. Except for *Perceptual Speed* and *General Information*, the tests were 25 minutes in length, a half hour including administration time. Perceptual Speed and General Information were always given together as they formed a one-hour unit. Otherwise, the order of administration of the tests was varied from college to college. All administrations took place in the fall of 1951.

The Criteria

The data used were found on the transcripts of the students' college records or on supplementary material provided by the colleges. Descriptions of the criteria follow.

Cumulative college average. This was the over-all college grade-point average. Many different marking systems were represented, but it was not necessary to convert all of these to a common scale, because separate correlation studies were undertaken for each college. The Cumulative Average was computed for all students who had completed at least a half year of work. Inclusion of students who did not finish college introduces an impurity into this criterion, because the grades are not earned in all years of college on the same basis. For example, it is somewhat easier to earn high grades in senior year than it is in freshman year. However, to have failed to include nongraduates in the cumulative average would have sharply reduced the number of cases in the study and might have eliminated the part of the range of test scores and grades that is of most interest to admissions officers.

Freshman grades. The freshman grade average was computed in the same way as the cumulative average. Freshman grade averages in specific course areas were also computed.

Major-field grades. The major-field grade was computed from the grades in the major-field courses taken at the participating college during junior and senior years. This criterion was computed for graduating students only. To simplify the tables given in this article and to increase the stability of the figures, the major fields were grouped into three groups: science and mathematics, social science, and humanities and languages. In all cases validity coefficients were computed separately for the individual major fields and were averaged by using *z* transformations and weighting by number of cases.[1]

Graduation-nongraduation. This was

[1] For graduating students, comprehensive examination grades in the major field were available for 6 out of 10 of the participating colleges. However, the validity patterns were found to be so much like those for the major fields that data on this criterion have been omitted from this article.

the simple dichotomy, graduation vs. nongraduation.

Reliability of the Criteria

For some of the colleges the grade average, or the major-field grades, or both, were computed separately by college year. This was done to provide a spot check on the estimated alternate-form reliability of these averages. To the extent that motivational or other factors change conditions during the course of the four years, the correlations are lowered. Therefore, the interyear correlations represent underestimations of the alternate-form reliability. All of the interyear correlations are based on graduating students, because relatively few of the nongraduates had a college record beyond freshman year.

The average of the interyear correlation figures for average grades (computed with z transformations) was .71. Since the separate single years can be considered to be alternate quarters of the criterion, it is appropriate to apply the Spearman-Brown formula to estimate the reliability of the four-year cumulative average. The corrected figure would be about .91. Furthermore, since the interyear correlations were computed for graduating students only, it is also reasonable to consider a correction for restriction of range in order to estimate the reliability of the cumulative average, which is used in this report for all students whether or not they graduated. The standard deviation of the cumulative average for graduating students was found to be on the average about 25% less than that of the cumulative average for all students. The correction for restriction of range would raise the reliability figure still farther. No attempt will be made to compute the exact correction, because corrections from .71 up to that level are subject to considerable distortion. It is clear, however, that the reliability of the cumulative average compares favorably with that of long, well-made aptitude tests. At the same time, it is well to remember that the "reliability" in some colleges might be partly a result of "halo." In addition, there are other aspects of grades such as promptness or neatness which may give them high consistency without necessarily reflecting consistent evaluation of achievement that is considered important.

For major-field grades, correlations between junior-year and senior-year grades were used. These were found to be .65, .75, and .71 for the three areas respectively. No correction for restriction of range is applicable. However, since the junior and senior years may be considered to be alternate halves of the major-field grade criterion, the Spearman-Brown formula may be used in arriving at estimations for the reliabilities of the two-year major-field average. The corrected figures were .79 for science and mathematics, .86 for social science, and .83 for humanities and languages. Since the best validities reported here or elsewhere do not approach the limit made possible by these reliabilities, the theoretical best possible prediction of college grades is still far away.

The Findings

Findings With Regard to Average Grades and Graduation

For the cumulative average, Table 1 summarizes for the ten colleges the validities of the SAT, High School Record, CEEB English Composition Test, and the experimental tests. To a considerable degree the validities are comparable from college to college. The only real exception to this is the large size of the validities of the experimental tests at College J. Comparisons among the validities of the tests are commented upon in a later paragraph when results from the several colleges are pooled.

TABLE 1

Summary Table of Validities for Cumulative Average

Variables	College									
	A₁	A₂	B	C	E	F	G	H	I	J
	$N =$ 449	$N =$ 172	$N =$ 154	$N =$ 870	$N =$ 222	$N =$ 246	$N =$ 118	$N =$ 481	$N =$ 579	$N =$ 190
SAT-V	.45	.56	.32	.44	.41	.41	.61	.39	.41	.43
SAT-M	.32	.21	.18	.31	.22	.27	.34	.22	.26	.27
High School Record	.39	.47			.62		.63	.42		
English Composition Test				.40					.30	
Social Studies Reading				.38		.35				.42
Science Reading	.28	.36	.25							
Inductive Reasoning			.23		.21					.49
Integration			.27	.24						.47
Sufficiency of Data	.32	.28								.45
Data Interpretation						.24	.37	.28	.30	
Visualization							.13	.10		
Best Arguments			.14			.13	.30	.12	.15	
Perceptual Speed	.17	.15		.18	.18				.07	
Carefulness				−.05	.00				−.06	
Picture Memory					.21	.16	−.02	.09		
Verbal Memory					.23	.25	.23	.18		
Number Memory					.16	.09	.15	.01		
Art Information	.27	.26		.28	.13				.23	
Literature Information	.36	.40		.37	.21				.25	
Social Work Information	.30	.18		.26	.28				.23	
Government Information	.40	.39		.41	.24				.35	
Biology Information	.27	.17		.22	.14				.09	
Physical Science Information	.23	.21		.21	.10				.20	
Mechanical Information	−.08	−.11		−.02	−.16				.01	

Table 2 shows the results pooled for all colleges; that is, averages across colleges have been computed. This makes convenient the comparisons among test validities for nine criteria: freshman average, cumulative average, graduation-nongraduation, and freshman and major-field average in each of the three areas.

Some of the experimental tests, particularly when their short length is considered, have substantial validities for cumulative average. A discussion of comparisons among the test validities, however, will be more appropriate in the next section where statistical corrections for restriction of range and for test length are applied.

Since freshman grades constitute part of the cumulative average, some similarity of validity coefficients for these two criteria is to be expected. However, the extreme closeness of the figures in the first

TABLE 2

Validities for Academic Criteria Averaged Over Colleges

Variables	Fresh. avg.	Cum. avg.	Grad.-nongrad.	Science & Math.		Social Science		Humanities & Lang.	
				Fresh.	Major	Fresh.	Major	Fresh.	Major
SAT-V	.44	.43	.09	.36	.35	.43	.43	.39	.34
SAT-M	.29	.27	.06	.37	.34	.20	.26	.18	.23
High School Record	.46	.46	.18	.41	.27	.34	.29	.37	.29
Social Studies Reading	.37	.38	.07	.35	.29	.40	.36	.31	.29
Science Reading	.29	.29	.04	.23	.21	.30	.31	.17	.29
Inductive Reasoning	.36	.31	.08	.38	.19	.24	.10	.39	.20
Integration	.32	.28	.09	.30	.28	.16	.17	.28	.22
Sufficiency of Data	.34	.34	.09	.42	.42	.25	.13	.21	.16
Data Interpretation	.25	.29	.12	.27	.18	.23	.29	.22	.21
Visualization	.17	.11	.01	.26	.19	.08	.04	.06	.02
Best Arguments	.14	.15	.04	.11	.08	.16	.12	.12	.16
Perceptual Speed	.17	.15	.03	.15	.06	.07	.12	.11	.26
Carefulness	−.04	−.05	−.02	−.02	−.04	−.03	.01	.00	−.08
Picture Memory	.12	.12	−.01	.13	.08	.14	.12	.05	.12
Verbal Memory	.21	.21	.09	.17	.07	.23	.24	.18	.09
Number Memory	.06	.08	.03	.07	.11	.04	.01	.01	.11
Art Information	.25	.25	.03	.18	.28	.27	.23	.22	.27
Literature Information	.35	.32	.09	.24	.26	.31	.31	.27	.27
Social Work Information	.24	.26	.06	.18	.24	.22	.24	.16	.28
Government Information	.37	.37	.11	.29	.34	.32	.33	.26	.24
Biology Information	.22	.19	.03	.23	.20	.19	.15	.10	.15
Physical Science Information	.20	.20	.05	.24	.27	.15	.12	.12	.18
Mechanical Information	−.02	−.04	−.02	.03	−.02	−.06	−.03	−.03	−.04

two columns of Table 2 shows that the tests that are valid for freshman grades are valid to much the same degree for upperclass grades. There is a very slight tendency for the cumulative average validities to be lower than the freshman validities, but the change is so slight as to be of no practical importance. The lack of substantial change in the size of the validity coefficients suggests that the factors favoring downward or upward changes listed earlier in this article either are not operative or approximately balance each other. These findings also support the viewpoint that for use in validity studies the freshman grade average is a satisfactory substitute for the four-year cumulative average.

The average validities for graduation-nongraduation are also given in Table 3. Apparently none of the tests in this study have an appreciable relationship to graduation-nongraduation, and high school record has very little. Before attempting to interpret this finding, it will help to look at the relationship of this criterion to grades.

The correlations between graduation-nongraduation and grades were found for the 10 colleges to range from .20 to .53. The weighted average is .44. However, these correlations are partly accounted for by an artifact of the situation. A check of the available data confirms what is, perhaps, a well known fact that, for those students who reach senior year, grade

TABLE 3

UNCORRECTED AND CORRECTED VALIDITIES FOR CUMULATIVE AVERAGE

Variables	College A_1 ($N = 449$)		College C ($N = 870$)		College I ($N = 579$)	
	Un-corrected	Corrected	Un-corrected	Corrected	Un-corrected	Corrected
SAT-V (90 min.)	.45	.54	.44	.54	.41	.58
SAT-M (60 min.)	.32	.41	.31	.40	.26	.45
SAT-V (25 min.)	.45	.51	.44	.51	.41	.55
SAT-M (25 min.)	.32	.39	.31	.38	.26	.42
Science Reading	.28	.37				
Soc. Stud. Reading			.38	.47		
Data Interpretation					.30	.46
Sufficiency of Data	.32	.42				
Integration			.24	.32		
Best Arguments					.15	.28
Literature Info.	.36	.53	.37	.57	.25	.51
Government Info.	.40	.60	.41	.62	.35	.62

averages are higher in senior year than in freshman year. By assuming that the students do not work any harder in their senior year, it can be argued, then, that the grading system changes; high grades are easier to get in senior year. Since most of the nongraduating students only received grades early in their college careers, when good grades were most difficult to earn, the correlation between grade average and graduation-nongraduation would almost certainly be higher than the actual relationship between graduation-nongraduation and scholastic success. One measure of this actual relationship is the correlation between freshman grades and graduation-nongraduation. In this study the correlation between graduation-nongraduation and cumulative average was found to be .46 at College B and .30 at College J, while the same figures for freshman average were only .25 and .15, respectively. These lower figures may be too low because of the elapse of time between freshman year and the time when many students withdraw, but they probably give a truer picture of the correlation between scholastic success and graduation-nongraduation than do the figures for cumulative average. This correlation is low, because so many things other than grades can cause a student to withdraw from college.

The implication of the still lower relationship between test scores and graduation-nongraduation seems to be that none of the colleges participating in this study admitted many students whose aptitude as measured by tests was so inadequate as to lead to either voluntary withdrawal or dismissal. This was true even for colleges E, G, and J, where the SAT statistics indicate that little or no selection occurred. On the other hand, to the small extent that grades do correlate with graduation-nongraduation, it may be said that withdrawal or dismissal occurs when students underachieve, that is, get grades which are lower than would be expected from their test scores. It is possible, of course, that the desire to leave college comes first for nonscholastic reasons and is followed by a drop in grades. In any case, the clear finding is that graduation-nongraduation does not serve as a predictable criterion against which it is possible to validate the kinds of tests tried out in this study.

Application of Statistical Corrections

In order to make appropriate comparisons between the validities of the tests, it is necessary to apply corrections for restriction in range on the SAT and for variations in the lengths of the tests. Unfortunately, it is often misleading to make statistical corrections, but it can be even more misleading to do without them. For this reason, figures for the reliability of the criteria have already been given both with and without corrections. Figures for the validities of the tests were given in Table 3 without corrections. Some of these validities will now be presented with corrections.

Corrections for restriction in range compensate for the high degree of selectivity employed by some of the participating colleges. The corrections alter the validity coefficients so as to equal the values which would have been obtained if the range of scores on which the validities were observed had been equal to that for the entire SAT candidate population on the date of the testing in March 1951. At this administration the standard deviation of SAT-V for the candidate population was 113, and that for SAT-M was 110.

Correction for test length was made by the Spearman-Brown formula after correction for restriction of range had been accomplished. The validity coefficients were corrected to simulate their value had every test been of "practical length," defined as 10 minutes for Perceptual Speed and 25 minutes for all others.

Either to average the validities and then apply two kinds of corrections or to apply the two corrections and then to average the corrected figures seemed to be covering up the observed validities with too much statistical folderol. Therefore, it was chosen not to do any averaging where corrections were made, but to select a few individual college findings with which to illustrate the effects of the proper corrections. The sample findings to be used for this purpose concern one criterion, the cumulative average, three colleges, A_1, C, and I, and a selection of variables including SAT, ECT, High School Record, and four experimental tests at each college. The colleges selected were the three largest except that College H was avoided, because only relatively unsuccessful tests were administered there (see Table 1). The tests selected for each college were those whose average validities for all *other* colleges were the highest. The only exception to this rule was a limit of two set upon the number of information tests selected. This selection technique, which was considered to be the equivalent of a cross-validation, led to the selection of the same two information tests for all three colleges: Literature Information and Government Information. The other tests were necessarily different at each college, since none of them was administered to more than one of the selected colleges.

Table 3 gives the selected data from Colleges A_1, C, and I. For each college the first column gives the observed correlations. The second column gives the same correlations after corrections were made for restriction in range on SAT-V and SAT-M and for test length.

It is apparent that, after the corrections are made on these data, neither the SAT nor the High School Record (College I) stand supreme as predictors. The highest validity is for Government Information. This probably reflects the importance to the criterion of width of serious reading outside of school requirements. The validity may be as high as it is because the student who abounds in this kind of information would probably possess both the aptitude measured by SAT-V and the willingness to spend time in serious extra study that produces a good high school record. While something more complex than breadth of information may be the desirable outcome of a college education, it is, nevertheless, undeniably true that the students who can demonstrate a wide knowledge of facts are often the ones who can think most clearly and are certainly the ones who fill up the academic honor roll. The runner-up tests, Literature Information, SAT-V, English Composition Test, and Social Studies Reading, all confirm the importance of serious reading to college grades.

Findings With Regard to Grades in Specific Areas

Table 2 compares the test validities for grades in three major-field areas with validities for freshman courses in these areas. As in the comparison between freshman average and cumulative average, there is evident here only a very slight drop in most validity coefficients between the freshman and upper-class years. In these tables the freshman and upper-class criteria do not overlap as they did in the case of freshman and cumulative average. The major-field grade criteria were averages of the appropriate course grades earned during the junior and senior years. In spite of this lack of overlap, the major-field validities have much similarity to the freshman validities. Here again the findings encourage the practice of using freshman grades as criteria of college success.

It is, perhaps, of interest to note that between freshman year and the upper-class years the validity of the High School Record for major-field grades falls off more sharply than do the validities of most of the test scores. It seems possible that the falling off of validities for High School Record in the case of specific course areas may be brought about by differences between general courses taken particularly in freshman year and the specialized, major-field courses taken during junior and senior years. The study techniques or other methods used to gain good grades in high school cannot be very different from those required in the more general college courses. However, quite different techniques may be required for specialized major-field work.

The validity of SAT-V is highest for social science; that for SAT-M is highest for science and mathematics. For humanities and languages the SAT-V validity is dominant over SAT-M as it is for social science, but both SAT validities are lower than they were for social science. The substantially lower validity of SAT for humanities and language grades cannot be attributed to low reliability of the criterion, because as shown in an earlier section, the reliabilities for the criteria in the three areas are, respectively, .79, .86, and .83.

The differences in validity for the SAT mentioned in the last paragraph are all significant at about the 1% level. Differences mentioned below in connection with the other measures are less significant. They are based on fewer cases. Even some relatively small differences will be mentioned to draw the reader's attention to differences of interest in judging whether further data are likely to reveal significant differences which will lead to useful differential prediction among the various specialized areas.

Among the experimental tests some different patterns of validity coefficients may be found for the three areas. Sufficiency of Data and Physical Science Information appear from these data to be superior to SAT-M as specific predictors of science and mathematics grades. While SAT-V is as good as any of the tests as a specific predictor of social science grades, Social Studies Reading and Data Interpretation are shown by these data to serve about as well. Government Information, as might be expected from the content of the test, has a relatively high correlation with social science considering that there has been no correction for its very short length, but it also has a surprising validity for science and mathematics. The only specific predictor for humanities and language grades is the poor general predictor, Perceptual Speed. A very good predictor of humanities and language grades, considering its short length, is also one that might be expected to be suitable for this purpose, Literature Information. However, this test shows no promise for differential prediction.

Summary

The College Board in 1951 initiated a validity study of the SAT and a group of experimental tests at 10 colleges. This article compares the validities of these tests for average freshman grades with their validities for the cumulative four-year average and graduation vs. nongraduation. The validities for freshman grades in certain subject-matter areas are compared with major-field grades in the same areas.

It was found that the pattern of test validities for the four-year criteria closely resemble those for the freshman criteria. These data show the high school record to be less good for predicting the quality of major-field work than it is for predicting freshman average grades. Tests of government and literature information were the most successful among the experimental tests. When corrections were made for restriction of range and for test length, these two tests were actually found to be more valid for predicting the cumulative four-year grade average than was the SAT. Neither the SAT nor any of the experimental tests had an appreciable validity for predicting graduation.

REFERENCES

1. Brush, E. N. Mechanical ability as a factor in engineering aptitude. *J. appl. Psychol.*, 1941, **25**, 300–312.
2. Dwyer, P. S., Horner, C., & Yoakum, C. S. A statistical summary of the records of students entering the University of Michigan as freshmen in the decade 1927–1936. Vol. 1, No. 4. *Administrative Studies*, Ann Arbor, Mich: Univer. Michigan, 1940.
3. Frederiksen, N. O. Princeton, N. J.: Educational Testing Service, Unpublished manuscript.
4. Guilford, J. P., & Lacy, J. I. (Eds.) Printed classification tests. Washington: U. S. Government Printing Office, 1947. (AAF Aviat. Psychol. Program Res. Rep. No. 5.)
5. Gulliksen, H. O. *Theory of Mental Tests.* New York: Wiley, 1950.
6. Stanford study of undergraduate education—A study of the variables used to determine freshman admission. Palo Alto, California: Stanford University, 1956. Unpublished manuscript.

Received September 13, 1957.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 2, 1958

A NOTE ON PART-WHOLE CORRELATION[1]

FREDERICK B. DAVIS

Hunter College

When a correlation coefficient is computed between total scores (such as the Performance scores derived from the Wechsler Adult Intelligence Scale) and part scores (such as the Object Assembly scores) that are included in the total scores, the resulting coefficient is spuriously high. There has been some confusion in the literature regarding the source and amount of this spuriousness. It is the purpose of this note to clarify the matter.

In deviation-score form of the original units of measurement, the product-moment correlation coefficient between scores on total t and on part a, which is wholly included in total t, may be written as follows:

$$r_{at} \equiv r_{(a)(a+b+\ldots+n)}$$

$$= \frac{\Sigma(a)(a + b + \ldots + n)}{\sqrt{\Sigma(a)^2}\,\sqrt{\Sigma(a + b + \ldots + n)^2}}.$$

Hence,

$$r_{at} = \frac{s_a + \sum_{b}^{n} s_j\, r_{aj}}{s_t}, \qquad [1]$$

where the subscript j denotes any part of total t except part a.

This coefficient is spuriously high because the errors of measurement in scores on part a are also in scores on total t. To obtain a coefficient free from this correlation of errors in common, a parallel form of part a, to be denoted Part A, may be employed. Part A is not, of course, included in total t. Then,

$$r_{At} = \frac{s_A r_{aA} + \sum_{b}^{n} s_j r_{Aj}}{s_t}, \qquad [2]$$

where r_{aA} is the reliability coefficient of part a.

Since parts a and A are parallel forms, $s_a = s_A$ and $r_{aj} = r_{Aj}$. Therefore, we may rewrite Equation [2] as:

$$r_{At} = \frac{s_a r_{aA} + \sum_{b}^{n} s_j r_{aj}}{s_t}. \qquad [3]$$

It is obvious from Equation [1] that

$$\sum_{b}^{n} s_j r_{aj} = s_t r_{at} - s_a.$$

Consequently, Equation [3] may be written as:

$$r_{At} = r_{at} + \frac{s_a(r_{aA} - 1)}{s_t}. \qquad [4]$$

Either Equation [3] or [4] may be used to compute the product-moment coefficient of correlation between a total and a part included wholly within the total if one wishes to report a coefficient free from the inflating effect of the correlation of errors of measurement common to both. Equation [4] will be the more convenient if r_{at} is known.

The difference between the values of Equations [1] and [3] or of Equations [1] and [4] may be written as:

$$r_{at} - r_{At} = \frac{s_a(1 - r_{aA})}{s_t}.$$

This difference, it should be noted, is not equal to the correlation of the errors of

[1] After this note had been accepted for publication, the writer's attention was called to a paper by Angoff (**1**) which he had not previously seen. Angoff's equation (**3**) and its derivatives constitute special cases of the writer's Equation [3] in the present paper.

measurement in scores on total t and on part a. That correlation coefficient, expressed in terms of the difference $r_{at} - r_{At}$, is:

$$r_{e_a e_t} = \frac{r_{at} - r_{At}}{\sqrt{1 - r_{aA}}\,\sqrt{1 - r_{tT}}}, \qquad [5]$$

where r_{aA} and r_{tT} are the reliability coefficients of part a and total t, respectively.

McNemar (**2**, p. 164) and other writers have referred to the correlation coefficient between a part score and the remainder of the total score as the part-whole correlation coefficient corrected for spuriousness. But it is obvious from its very definition that such a coefficient is really not a part-whole correlation coefficient; it is instead a part-remainder correlation coefficient, it needs no correction for spuriousness, and it may be denoted and computed as follows:

$$r_{(a)(t-a)} = \frac{s_t r_{at} - s_a}{\sqrt{s_a^2 + s_t^2 - 2 s_a s_t r_{at}}}. \qquad [6]$$

Use of Equations [4], [5], and [6] may be illustrated with data pertaining to the relationship between Part 11 (Object Assembly) and the Performance total score (the sum of Parts 7, 8, 9, 10, and 11) of the Wechsler Adult Intelligence Scale (**3**). The basic data, reported in terms of Wechsler's Scaled Score units for a group of 200 eighteen–nineteen year olds, are as follows (when X_P denotes a Scaled Score on the Performance total and X_O a Scaled Score on the Object Assembly part):

$$\begin{array}{ll} \overline{X}_O = 10.00 & r_{OP} = .82 \\ s_O = 2.79 & r_{oO} = .65 \\ \overline{X}_P = 49.43 & r_{pP} = .93 \\ s_P = 11.83 & \end{array}$$

The correlation coefficient of .82 between scores on Part 11 and the Performance total (r_{OP}) was computed directly from the data. Equation [4] yields a value of .74 for the correlation between scores on Part 11 and the Performance total free from the spurious inflation owing to the perfect correlation of errors of measurement common to both scores. Equation [6] yields a value of .71 for the correlation between scores on Part 11 and the sum of the remaining parts of the Performance total. Equation [5] yields a value of .52 for the correlation between errors of measurement in the entire Performance total and in Part 11 alone.

As would be expected, the coefficients yielded by Equations [1], [3] or [4], and [6] range themselves in order of decreasing magnitude. The coefficient of .82 indicates the actual relationship of two partially overlapping variables—scores on Part 11 and the Performance total in the sample of 18–19 year olds. On the other hand, the coefficient of .74 indicates the relationship of two entirely separate variables that measure the same abilities (plus chance) as Part 11 and the Performance total in the same sample of 18–19 year olds. This is a part-whole coefficient properly corrected for spuriousness owing to the correlation of errors in common. The coefficient of .71 indicates the actual relationship of scores on Part 11 and the sums of scores on other parts included in the Performance total. This is a part-remainder coefficient.

Of the three coefficients, the one having the value of .74 is most meaningful for comparison with the great majority of intercorrelations reported among mental tests. This is because such intercorrelations are ordinarily based on separate tests and are not inflated by correlation of errors of measurement in common. The coefficient of .82 is of fundamental utility in computing variances, standard errors, etc. The meaning of the coefficient of .71 is clear, but this type of coefficient is not commonly of practical utility. Each of these coefficients has its own particular merit and the distinctions among them should be recognized so that one will not be confused with another.

REFERENCES

1. ANGOFF, W. H. A note on the estimation of nonspurious correlations. *Psychometrika*, 1956, **21,** 295–297.
2. MCNEMAR, Q. *Psychological statistics.* New York: Wiley, 1955.
3. WECHSLER, D. *Manual for the Wechsler Adult Intelligence Scale.* New York: Psychological Corp., 1955, Tables 6, 7, and 10.

Received November 14, 1957.

Journal of Educational Psychology
Vol. 49, No. 2, 1958

THE INFLUENCE OF CONSISTENT AND INCONSISTENT GUIDANCE ON HUMAN LEARNING AND TRANSFER[1]

BERNARD M. ARONOV

University of Florida[2]

In 1928, Goodenough (**7**) reported as a finding of her study on anger in young children an apparent relationship between inconsistency of parental discipline and frequency of anger outbursts. Other studies on the effects of consistency and inconsistency followed. These dealt with various ways in which consistency or inconsistency is expressed, as for instance in parental demands, commands, etc. (**1, 2, 4, 5, 8, 12**). The results of these studies all suggested the conclusion that inconsistency in the behavior of an authority figure toward a child has disturbing effects both on the child's immediate behavior and on his subsequent personality development. Support for this conclusion came from animal studies in which random reinforcement was a variable (**11, 15, 16**). The last study on the effects of these variables appeared in 1952 (**8**), and our textbooks speak of the detrimental effects of inconsistency as established fact (e.g., **3, 6, 9, 10, 13**).

The study here reported arose, however, as a result of an impression that the work done on the problem does not justify the conviction shown with regard to the detrimental effects of inconsistency. For, while the data strongly support the contention, e.g., that parental inconsistency has damaging effects on a child's behavior, the nature of the studies gives reason to question the validity of the data. The direct data on inconsistency come from case history and observational studies, studies too loosely designed to control for the possibility that it is the person being inconsistent rather than the inconsistency itself which is causing the damage. Baldwin, et al. (**1**), for instance, found that inconsistency *figured in* the rejecting parent's behavior in that discipline, decisions, etc., were based on the parent's convenience. This finding would suggest that inconsistency is one avenue through which *rejection* is expressed, but for which the inconsistency could be irrelevant. The studies involving random reinforcement seem better controlled, but only one (**15**) includes a study of important transfer effects, and in every case we have no way of knowing how far we can generalize from infrahuman to human *S*s. In brief, it appears that we actually do not know that inconsistency itself has a detrimental effect on behavior.

The intent of this study, then, was to isolate and study the variables of consistency and inconsistency in a controlled laboratory setting. The study was not designed to investigate the effects of parental consistency or inconsistency. Although primary interest has been in the effects of parent-child inconsistency, it was considered important to test the specific effects of these variables apart from other conditions. An attempt was made simply to answer the following question: If *S*s are given consistent or inconsistent guidance while learning to solve a maze problem, in what ways will their learning behavior be affected both in the immediate learning situation and later when they are

[1] This paper is a condensation of the author's doctoral dissertation, completed in 1956 at the University of Florida under the direction of Rolland H. Waters, who was also kind enough to read and make valuable suggestions about the present paper. The writer is also grateful to Henry S. Curtis, Morton S. Slobin, and Stanley Spiegel for their help with this paper.

[2] Now at the VA Regional Office, Cleveland, Ohio.

no longer being guided and are confronted with a similar but different problem to solve?

Procedure

Eighty-eight college students, male and female, under 26 years old, and essentially inexperienced with maze problems, served as *S*s. As each *S* appeared at the experimental room, he or she was assigned randomly to one of three groups, known as Groups I, II, and C. The *S*s were seated before a shield which obscured the apparatus and the experimenter. They were given a general description of the type of maze they were to learn, and told that their purpose was to learn to guide a stylus from start to goal without error. They were told further that when they reached the goal at the end of each run both the red and the green light suspended before them would flash to signal the end of a trial. In addition to this general orientation, *S*s in Groups I and II were told that when they made a correct turn the green light would flash, and when they made a wrong turn the red light would flash.

The stylus was placed in the *S*s hand and guided to the starting point of a standard 10-turn Warden **U**-type maze (**14**) employed, and the *S* was told to start. Light cues were given Group I *S*s consistently according to instructions. Unknown to Group II *S*s, however, the light cues given them were wrong at three of the ten choice points on each trial. Also, the choice points at which wrong cues were given were varied from trial to trial according to a prearranged pattern. Group C *S*s were given no guidance and served as the control group.

Whether or not they reached the criterion of one errorless run, all *S*s were required to run trials, after which all were stopped and transferred to the lateral reverse of the practice maze pattern. Here they were told that their task and purpose were the same, but that the only light cues they would see would be those at the end of each run. All *S*s were then allowed to run until they reached the criterion of one errorless trial. Records were kept of errors and time per trial, and of trials to criterion, and notes were taken of spontaneous behavior exhibited. Following completion of the second maze problem, *S*s were interviewed with regard to their impressions of the experimental experience.

Results

The quantitative results are summarized in Tables 1 and 2. It will be noted that no figures are given for trials to criterion on the practice maze. The reason for this is that since only 11 Group I *S*s and eight Group C *S*s reached criterion in 15 trials, it was not possible to compute mean trials to criterion. Instead, the percentages of *S*s in each group who reached criterion were computed, and these percentages were compared using the chi-square method.

For the inconsistently guided Group II,

TABLE 1
Practice Maze Performance

	Means and Standard Deviations			
Group	Errors		Time	
	M	*SD*	M	*SD*
I	51.77	12.43	425.30	157.83
II	86.43	13.22	668.57	324.67
C	56.50	11.66	523.70	212.54
	F Ratios and *t*s			
	Errors		Time	
	F	*t*	*F*	*t*
I—II	1.13	10.22**	4.23*	2.76**
I—C	1.14	1.60	1.81	2.14*
II—C	1.28	9.11**	2.33	2.91**

* Significance at .05 level.
** Significance at .01 level.

TABLE 2

Transfer Maze Performance

Group	Means and Standard Deviations					
	Errors		Time		Trials	
	M	*SD*	M	*SD*	M	*SD*
I	48.10	30.72	368.20	212.27	16.06	7.46
II	109.14	68.19	663.86	482.82	32.75	25.61
C	57.00	47.98	403.93	243.75	19.40	15.16
	F Ratios and *t*s					
	Errors		Time		Trials	
	F	*t*	*F*	*t*	*F*	*t*
I—II	4.92*	4.34**	5.17*	2.98**	11.89*	3.32**
I—C	2.44*	0.86	1.31	0.63	4.14*	1.14
II—C	2.02	3.34**	3.92*	2.56*	2.85*	2.39*

* Significance at .05 level.

** Significance at .01 level.

mean total errors and time were significantly greater than those for Groups I and C, both on practice and transfer mazes. A significantly smaller proportion of Group II *S*s reached the criterion within 15 trials on the practice maze (chi square 14.05, significant beyond the .01 level). Group II *S*s required significantly more trials to reach criterion on the transfer maze than did Groups I and C. Group C practice maze time was significantly greater than that of Group I, but otherwise Group I performed only slightly and insignificantly better than did Group C.

While variances did not differ significantly for the practice maze, they did for the transfer maze. Group II variances were significantly greater than those of Group I for all measures, and greater than those of Group C for time and trials to criterion. Also, Group C variances were significantly greater than those of Group I for errors and trials to criterion.

The behavioral data characterized Group I *S*s as initially dependent upon the light cues but as gradually showing less dependence upon them. Group II *S*s were more characteristically confused by the lights at first and then reacted to them in one of three ways: either they rebelled against instructions and ignored the lights, they were confused and ambivalent about them, or they followed them passively. Group II *S*s tended also to be uneasy about verbalizing doubts concerning the accuracy and usefulness of the light cues; those bold enough to rebel against the use of the cues were quite outspoken, but at the other extreme those who followed the cues passively distorted their perceptions of the situation so far as to insist that the cues were helpful. Further, transfer maze performances for Group II were related to the degree to which *S*s had ignored the light cues on the practice maze, i.e., those who ignored the lights tended to do as well as the best in Groups I and C, etc. Group C *S*s approached the mazes in a matter-of-fact, business-like manner.

Discussion

The excessive variance of Group II transfer maze performance, together with the protocol material, makes an interpretation of the effect of the inconsistent guidance difficult. On the one hand, the group performances suggest that to a significant degree Group II was adversely affected by the inconsistent guidance. On the other hand, however, the magnitude of Group II variance cautions against drawing such a broad conclusion, because the inconsistency can hardly be said to have had a very uniform effect on Group II *S*s.

The results seem to become understandable when Group II variance and behavioral data are considered in detail. First, the possibility of a sampling bias contributing to the variability can be discarded on the grounds that the groups showed similar variability on the practice maze. Can it be concluded then that the inconsistency itself produced the variability, or did the inconsistency bring into play personal variables which determined individual performances?

The behavioral data suggest the latter of the two possibilities. It appears that the inconsistency provoked three grossly different personal reactions, i.e., a defiant and rebellious one, a confused and ambivalent one, and a passive one. It appears further that the particular personal reaction provoked was related to transfer maze performance. It is true that the inconsistent cues had the initial effect of confusing all the Group II *S*s (perhaps accounting for the more similar variance of practice maze performance), but this effect did not last for those *S*s who were able to break away from the light cues and to attend to cues from the maze itself. Lasting confusion and damage to performance seemed to occur primarily when *S*s could not break away from the inconsistent guidance. These *S*s emerged from the practice maze experience with little useful information to apply in dealing with the transfer maze. It would seem necessary to conclude, then, that the inconsistent guidance had the immediate effect of confusing the recipient, but that its effects were temporary unless the recipient was unable to rebel against the inconsistent guidance.

Some clues are present which suggest an explanation for this behavior of Group II *S*s. It appears that the more ambivalent and passive *S*s were those who seemed to fear offending the experimenter and/or being embarrassed by questioning the cues. For personal reasons these people seemed to feel uncertain enough in the relationships with themselves and/or with the experimenter to feel that it was important not to question the experimenter too seriously if at all—the safest reaction being complete passivity.

Finally, a few words might be said about the variance differences between Groups I and C, where actual performances did not differ significantly. It is felt that the guidance given Group I encouraged group conformity of performance, while no guidance perhaps allowed Group C *S*s to develop whatever potentials they had.

Implications

If the results of this study are dependable, they raise important questions about the origins of behavior pathology. Broadly speaking, the results make it difficult to maintain the position that a particular type of experience will affect personality in a particular way. We are confronted again by that constant source of irritation, the intervening variable. In this particular instance, the effect that the "experience" had was apparently influenced by how the *S* perceived the situation, and that perception seemed in turn influenced by how secure the *S* felt in relation to himself and/or to the experimenter. What apparently was important here was whether the *S* perceived the situation as one in which he could comfortably question the misguiding information he was receiving.

It would seem important, then, in understanding the origins of behavior disturbance, to study some of the intervening variables which could play a role in determining the effect that a particular experience might have on the developing personality. With regard to the results of the present study, it would seem important to know more about the variables which influence a person to perceive a situation as one in which he could or could not comfortably question inconsistent guidance being given him by an authority figure. In the final analysis, a study of such intervening variables may reveal that what a parent actually does or does not do with regard to his child is not nearly so important for the developing personality as is, for instance, the interpersonal relationship in which this act occurs.

Summary and Conclusions

This study was designed to answer the question: If *S*s are given consistent or inconsistent guidance while learning an initial maze problem, in what ways will their learning behavior be affected both in the immediate and in a transfer situation? On a 10-turn Warden **U**-type maze, *S*s were given either consistent, inconsistent, or no guidance. After 15 trials under one of these conditions, all *S*s were transferred to the lateral reverse of the initial maze where all were required to run without guidance until one errorless run was achieved. After learning the transfer maze, *S*s were interviewed for impressions of the experiment. The results suggested the following statements in answer to the motivating question:

1. The influence of consistent guidance is not markedly different from that of no guidance.

2. While inconsistent guidance is being given it has a confusing and generally detrimental influence on learning as compared with the influence of consistent or no guidance.

3. Inconsistent guidance does not necessarily have lasting damaging influence on learning behavior.

4. Lasting damage to learning behavior results from inconsistent guidance when the recipient of the guidance is for some reason unable to rebel and ignore the guidance.

References

1. Baldwin, A. L., Kalhorn, J., & Breese, F. H. Patterns of parent behavior. *Psychol. Monogr.*, 1945, **58**, No. 3 (Whole No. 268).
2. Baruch, Dorothy W. A study of reported tension in interpersonal relationships as co-existant with behavior adjustment in young children. *J. exp. Educ.*, 1937, **6**, 187–204.
3. Brooks, F. D. *Child phychology*. Boston: Houghton Mifflin, 1937.
4. Bühler, Charlotte. Clinical stuides of mother-child relationships. *Psychol. Bull.*, 1940, **37**, 586. (Abstract)
5. Friedlander, Dorothea. Personality development of twenty-seven children who became psychotic. *J. abnorm. soc. Psychol.*, 1945, **40**, 330–335.
6. Garrett, H. E. *Great experiments in psychology*. New York: D. Appleton-Century, 1941.
7. Goodenough, Florence L. *Anger in young children*. Inst. of Child Welf. Monogr. Series. Minneapolis: Univer. of Minnesota Press, 1931.
8. Havighurst, R. J. The functions of successful discipline. *Understanding the child,* 1952, **21**, 35–38.
9. Linton, R. *The cultural background of personality*. New York: Appleton-Century-Crofts, 1945.
10. Louttit, C. N. *Clinical psychology*. New York: Harper, 1947.
11. Maier, N. R. F. *Frustration, the study of behavior without a goal*. New York: McGraw Hill, 1949.
12. Meyers, C. E. An experimental study of the effect of conflicting authority upon child behavior. *Psychol. Bull.*, 1941, **38**, 710. (Abstract)

13. Shaffer, L. F. *The psychology of adjustment.* Boston: Houghton-Mifflin, 1936.

14. Warden, C. J. Primacy and recency as factors in cul-de-sac elimination in a stylus maze. *J. exp. Psychol.,* 1924, **7,** 98–116.

15. Wike, E. L. Extinction of a partially and continuously reinforced response with and without a reward alternative. *J. exp. Psychol.,* 1953, **46,** 255–260.

16. Wilcoxon, H. C. "Abnormal Fixation" and learning. *J. exp. Psychol.,* 1952, **44,** 324–333.

Received November 18, 1957.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 2, 1958

A TECHNIQUE FOR MEASURING CLASSROOM BEHAVIOR

DONALD M. MEDLEY AND HAROLD E. MITZEL

Division of Teacher Education of the Municipal Colleges of New York City

One of the most difficult problems that must be solved before useful results can come from research into the relationship between teacher personality and pupil growth is that of securing objective measures of the teacher's personality as it functions in the classroom. The usual approach to this problem has been to use ratings by supervisors or specially trained observers, but, despite all attempts to improve them, such ratings are still biased, subjective, and in many cases uninterpretable by anyone, even the rater himself.

Whatever value such ratings have arises from the fact that they are based on observations of the teacher while he is teaching; their most serious limitations arise from the fact that the evaluative judgment of the rater intervenes between the behavior and the score supposed to reflect it. There are at least two sources of variation introduced here that attenuate the validity of the ratings by distorting measured differences between teachers. The cues upon which the observer bases his judgment and the relative weights assigned to them are both allowed to vary from observer to observer to some unknown degree. By providing a schedule for recording behaviors listing the cues to be responded to, the first source of error may be virtually eliminated. By making the assignment of weights a clerical task done by someone other than the observer, the second may also be made negligible.

As a part of a longitudinal study of graduates of the Teacher Education program of the municipal colleges of New York City (City, Hunter, Brooklyn, and Queens) carried out in the Office of Research and Evaluation of the Division of Teacher Education, a technique for objectively observing and recording classroom behaviors was developed. The Observation Schedule and Record (OScAR) was constructed by modifying and combining the methods proposed by Cornell (**1**) and Withall (**4**) on the basis of the results of tryouts of the two techniques. Three basic changes were made.

Inspection of the reliabilities of the scales prepared by Withall and Cornell showed that some of them suffered from a lack of observer agreement to a degree that seriously impaired their accuracy (**2**). Accordingly, the first change was designed to increase observer accuracy. If an observational technique is such that it takes a highly trained observer to use it successfully, it has limited usefulness, and results of future measurements may be suspect because the observers may be inadequately trained. For this reason, the scales of both Cornell and Withall were redefined in somewhat simpler terms for use in the OScAR in order to minimize the amount of training necessary for its use.

Experience with these two techniques also showed that the often-adopted practice of sending several observers into the classroom together (presumably so that one observer can record what another misses) is uneconomical. A score based on observations made by two observers who see a teacher at different times is actually more reliable than one based on observations made by two observers who see the teacher at the same time; and it seems intuitively obvious that the former score is more valid as well, since the behavior sample obtained is twice as great. The OScAR was therefore designed to be used by a single observer visiting a classroom by himself.

The third change involved was the separation of the process of scoring from the process of observing teacher behaviors. The OScAR was designed to permit the recording of as many aspects of what goes on in a classroom as possible, regardless of their relationship to any dimension or scale. The observer's sole concern was to see and hear as much of what was going on as he could, and to record as much of it as the structure of the OScAR permits, without any attempt to evaluate what he saw.

Description of the OScAR Technique

The OScAR technique is both a method of observing and a method of recording classroom behavior; in the interests of simplicity the two aspects will be described simultaneously.

The observer making a visit to a classroom arrives at—or near—a prescheduled time, so it is usually not necessary for him to greet the teacher or class when he arrives. Instead, he tries to enter and take a seat at the back of the room as unobtrusively as possible. He first notes the time and the number of pupils present in the spaces at the upper left corner of the "front" of a specially printed 5 × 8 card (see Fig. 1).[1] Then he starts his stopwatch and begins to record behaviors on the front of the card by checking as many of the items in the Activity Section as describe what he sees.

The Activity Section consists of 44 activities likely to be observed in a classroom, such as "teacher works with individual pupil," "pupil writes or manipulates at his seat," "pupil laughs." Varying numbers of the Activity items may be checked, according to how many different kinds of activities are going on at one time.

The observer then concentrates on the Grouping Section. The Grouping Section lists four sizes of groups from "at least half of class in group with teacher" and "at least half of class in group without teacher" to "pupil as individual." In Column I he checks each type of administrative group (i.e. group apparently set up by the teacher) that he can detect in the class and each type of *social* group he observes—a social group being defined as one in which there is pupil-pupil or pupil-teacher interaction.

Next the observer checks the type of instructional materials being used, in the Materials Section, which lists various learning aids and materials such as blackboard, audio aid, text or workbook. All through this initial period, the observer keeps alert for any type of activity, grouping, or material not already checked, and checks the appropriate item for each one as it occurs. No item on this side of the card is checked more than once during this time, however. Items in the Signs Section (which consists of items considered symptomatic of classroom climate, like "teacher shows affection for pupil" and "pupil moves freely") are marked with a plus sign if and when they are observed. At the end of five minutes the observer briefly considers each item in this section not already marked, and marks it either plus or zero.

As soon as he has done this, the observer stops his watch and turns the card over (See Fig. 2). In the Subject Section, which lists the 10 most common subject areas, he checks in Column I whichever of the 10 areas of instructional activities has received most attention during the five minutes just ended.

The observer then starts his stopwatch again and begins to tally each statement the teacher makes in one of five categories: Pupil-Supportive, Problem-Struc-

[1] Tables A through G and Figures 1 and 2 have been deposited with the American Documentation Institute. Order Document No. 5556, remitting $1.75 for 35 mm. microfilm or $2.50 for 6 by 8 in. photocopies. Typescript copies of a more detailed version of this paper containing all tables will be furnished on request to the authors while the supply lasts.

turing, Miscellaneous, Directive, Reproving. He makes a tally in Column II of the Expressive Behavior Section in the line corresponding to the category in which each statement is classified.

At the same time, he watches for changes of expression on the teacher's face, such as smiles, frowns, and scowls, and for expressive gestures such as nods, threatening glances, and body movements. Each time he observes a look or gesture which he judges to express approval of or affection for a pupil, the observer makes a tally in Column II after Item K1; each time he observes a look or gesture which he judges to be hostile or reproving, he makes a tally after K7.

This continues for a second period of five minutes. At the end the observer stops his watch again and fills out Column II in the Subject Section just as he filled out Column I at the end of the first five-minute period. He then turns the card over, starts his stopwatch again, and proceeds as in the first period for five minutes more, except that he uses Column III rather than Column I. This alternation of sides of the card is continued until six five-minute periods of observations are completed.

Collection of Data

The observations which form the primary data of this study were made with OScAR in the classrooms of 49 beginning teachers in public elementary schools in New York City over a period of approximately 10 weeks. Of the 49 teachers, 46 were female, 3 male. The teachers were scattered among 19 schools in four boroughs, the number of teachers in a single school ranging from two to five. Twenty-three of the teachers taught Grade 3, thirteen Grade 4, nine Grade 5, and four Grade 6.

Observers worked in pairs, two observers visiting a school together. In most cases, all of the teachers in a school were seen by both observers in a pair on the some day, although in no case did two observers visit the same teacher at the same time. No attempt was made to control the type of activity observed; all that was asked was that the teacher and the class be present in the classroom.

A number of minor shortcomings in the original OScAR (**2**) having been noticed, it was revised to the form described in this report. The new form was adopted at the beginning of the second round of visits. The pairs of observers who went to the schools together were reshuffled somewhat, and a new schedule in which each observer was to see each teacher once again was set up. The first visits were made on January 24, 1955, and the last on Tuesday, April 5, 1955.

Analysis of the Observational Records

The analysis of the data followed four steps. First, a preliminary study was made of each item to find out whether there were reliable differences in the number of times the behavior was observed in the classrooms of different teachers. Next, the items were combined into 14 "keys," which were scored. Third, a factor analysis of scores on these 14 keys was made; and finally, the keys were combined into three factor dimensions.

The results of the analysis of individual items are given in Tables A through F.[1] Except in the case of a few items that were highly reliable by themselves, those items that discriminated well were combined into provisional keys on the basis of a priori judgment that they belonged together.

For example, the following three items from the Activity Section:

E1. pupil talks to a group
E5. pupil demonstrates or illustrates
E10. pupil leads the class

were combined into a single key called "Pupil Leadership Activities." (The com-

TABLE 1
INTERCORRELATIONS AND RELIABILITIES (IN THE DIAGONAL) OF SCORES ON FOURTEEN KEYS FOR OSCAR
($N = 49$)

Scoring Key	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
(1) Time spent on reading	(.641)	.226	.543	−.296	.196	−.101	.187	−.593	−.067	.612	.518	.315	.017	.080
(2) Problem-structuring teacher statements		(.801)	.137	.002	−.157	−.294	.244	−.225	−.405	.126	.460	.656	.057	−.275
(3) Autonomous administrative groupings			(.632)	.016	.270	−.163	.026	−.189	−.098	.454	.159	.039	.331	.489
(4) Pupil leadership activities				(.717)	.005	−.423	.393	.196	−.374	−.350	.123	−.285	.273	.017
(5) Freedom of movement					(.625)	−.059	.075	−.216	.198	.152	−.115	−.108	.302	.378
(6) Manifest teacher hostility						(.916)	−.409	−.034	.626	−.136	−.348	.168	−.134	−.013
(7) Supportive teacher behavior							(.841)	−.381	−.417	.023	.327	.209	.110	−.201
(8) Time spent on social studies								(.718)	.106	−.267	−.337	−.405	−.062	.064
(9) Disorderly pupil behavior									(.891)	−.117	−.288	−.200	.144	.418
(10) Verbal activities										(.617)	.299	.253	−.124	.186
(11) Traditional pupil activities											(.698)	.382	−.083	−.187
(12) Teacher's verbal output												(.796)	−.226	−.434
(13) Audio-visual materials													(.739)	.435
(14) Autonomous social groupings														(.605)

TABLE 2

Loadings of Fourteen OScAR Scoring Keys on Three Orthogonal Factors
($N = 49$)

Scales	Emotional Climate	Verbal Emphasis	Social Structure	Specific Factors	Communalities
(1) Time spent on reading	+.09	+.85	−.03	.52	.73
(2) Problem-structuring teacher statements	+.54	+.31	−.43	+.66	.57
(3) Autonomous administrative groupings	+.15	+.59	+.49	+.63	.60
(4) Pupil leadership activities	+.55	−.41	+.30	+.66	.56
(5) Freedom of movement	−.05	+.29	+.48	+.83	.32
(6) Manifest teacher hostility	−.76	+.05	−.22	+.61	.62
(7) Supportive teacher behavior	+.63	+.11	−.02	+.77	.41
(8) Time spent on social studies	−.16	−.60	+.13	+.78	.40
(9) Disorderly pupil behavior	−.73	+.06	+.20	+.65	.58
(10) Verbal activities	+.03	+.65	+.01	+.76	.42
(11) Traditional pupil activities	+.47	+.43	−.32	+.70	.51
(12) Teacher's verbal output	+.15	+.45	−.60	+.64	.59
(13) Audio-visual materials	+.18	+.05	+.55	+.82	.34
(14) Autonomous social groupings	−.22	+.18	+.72	+.63	.60

TABLE 3

Intercorrelations Among Three Factor Scales Based on OScARs of 49 Beginning Teachers
(Reliabilities in The Diagonal)

Scale	EC	VE	SS
Emotional Climate	(.903)	−.004	−.110
Verbal Emphasis		(.770)	+.028
Social Structure			(.826)

position of each of the 14 keys found to discriminate is given in Table G.) The reliability of each key was estimated from a three-way analysis of variance under mixed-model assumptions—teachers and visits being regarded as random effects and items as a fixed effect.

The coefficient of reliability so obtained is a maximum likelihood estimate of the expected correlation between the mean of all the scores assigned to the teachers by the six observers on the basis of the twelve visits made, and means of scores that would be assigned to the same teachers by six different observers visiting their classrooms at twelve other times. Errors arising from three potentially important sources are taken into account: errors resulting from fluctuations in teacher and pupil behaviors during several weeks, errors resulting from differences in ways in which various observers would tally identical behaviors, and errors resulting from the failure of an observer to note and record all that happens during a five-minute period.

Table 1 shows the reliabilities of all 14 scoring keys and their intercorrelations. The sizes of the reliability coefficients indicate that these teachers' classes differed widely with respect to what was going on in them.

The intercorrelations among the 14 dimensions suggest that the differences might be described in terms of fewer than 14 variables, so a centroid factor analysis was made and three factors extracted. The centroid factor matrix was rotated orthogonally twice according to the procedure proposed by Reyburn and Taylor. Table 2 shows the loadings of the original keys on the three factors after rotation. The factors were named Emotional Climate, Verbal Emphasis and Social Structure.

Three scales were constructed by combining, with equal weights and the signs indicated by the loadings, the scores on those keys most highly loaded on each factor. Table 3 shows the reliabilities of these three scales and their intercorrelations. The three scales are practically independent of one another (as would be expected) and are highly reliable.

Discussion

The effort made in this study to secure quantitative, objective information about happenings in ordinary classrooms and typical learning situations was not intended to imply that ratings by supervisors and other qualified observers may not serve a useful purpose. It arose from the conviction that there are purposes such ratings cannot serve. One such purpose is research into the nature of teacher effectiveness—research seeking to answer questions about how teachers influence pupil learning.

Information that effective teachers are warm and friendly, or firm but fair, or that they explain things clearly, is useful in this sense only if these terms are operationally defined. If such operational definitions must be phrased in terms of expert judgment, they can tell us only about expert judgment. Whatever inferences research with a technique such as OScAR justifies will tell educators what a teacher should do in specific terms—not what someone's reaction to his behavior ought be.

A study of the factorial structure of the 14 scoring keys indicates that the OScAR technique gives reliable information about three relatively discrete dimensions of classroom behavior—the social-emotional climate, the relative emphasis on verbal learnings, and the degree to which the social structure centers about the teacher. Certainly there must be many other important differences in ways teachers and pupils behave that are not included in this list. It is important that such differences be identified and techniques developed for observing them.

The potential importance of the kind of objective data about classroom behavior that can be obtained in this way is very great. Practical problems such as how to select students likely to become successful teachers, how to screen out those who cannot get along with children, and what ought to be the content of teacher training, can be solved in no other way than by studying teachers' classroom behavior.

Summary and Conclusions

The OScAR was developed as a device for securing a record of behaviors of teachers and pupils observed by a classroom visitor. It was used in a series of 588 half-hour visits made by six observers visiting 49 teachers twice each. Items which on the basis of content appeared to belong together were grouped into 14 keys which were found to have reliabilities of at least .60. A factor analysis identified three orthogonal factors accounting for most of the observed differences.

The three aspects in which the behaviors observed in the 49 classrooms differed were: Emotional Climate, having to do with the relative amount of hostility observed; Verbal Emphasis, having to do with relative emphasis on verbal and traditional schoolroom activities; and Social Structure, having to do with the relative degree of pupil-initiated activity. These three aspects were found to be orthogonal —a hostile class was no more likely to be verbal, or to have a restricted social organization than one less hostile.

It was concluded that (*a*) relatively untrained observers using an instrument like OScAR can develop reliable information about differences in classrooms of different teachers, (*b*) that the OScAR technique is sensitive to only three of many dimensions that probably exist, and (*c*) that observations made with instruments of this type can contribute to the solution of many important problems having to do with the nature of effective teaching.

REFERENCES

1. CORNELL, F. G., LINDVALL, C. M., & SAUPE, J. L. *An exploratory measurement of individualities of schools and classrooms.* Bureau of Educ. Res., Coll. of Educ., Univer. of Illinois, September, 1952.

2. MEDLEY, D. M. & MITZEL, H. E. *Studies of teacher behavior: The refinement of two techniques for assessing teachers' classroom behaviors.* Office of Research and Evaluation, Div. of Teacher Educ., Bd. of Higher Educ. of the City of New York, October 1955.

3. THOMSON, GODFREY. *The factorial analysis of human ability.* New York: Houghton Mifflin, 1951

4. WITHALL, J. G. The development of a technique for the measurement of social-emotional climate in classrooms. *J. exp. Educ.*, 1949, **17,** 347–361.

Received December 12, 1957.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 2, 1958

A FAILURE IN THE PREDICTION OF PUPIL-TEACHER RAPPORT[1]

WILLIAM RABINOWITZ AND IRA ROSENBAUM

Division of Teacher Education, Municipal Colleges of New York City

The essential purpose of this study was to determine the success with which several test instruments could predict the pupil-teacher rapport achieved by a group of teachers. The participating subjects took the tests as student-teachers; the criterion measure of rapport was obtained approximately one year later in the classrooms of the same subjects, who were then completing their first year of teaching. By employing test and criterion measures that were clearly separated in time, the study attempted to determine the predictive validities of the tests for the criterion used.

Pupil-teacher rapport was measured through pupil responses to questions about their class and their teacher. The variable to be predicted was, therefore, not teacher behavior, but pupil reactions to teacher behavior. Since it cannot be assumed that pupils respond in similar fashion to similar teacher behaviors, tests that validly predict various aspects of the classroom behavior of teachers might not predict pupil responses to such behavior. For this reason, a number of measures based on the teachers' classroom behavior were included in the study as a "bridge" between the test measures and criterion measure of major interest.

METHOD

During the 1953–54 academic year, over 1600 students who were enrolled in student teaching in the four municipal colleges of New York City were given a battery of tests. Some of the instruments of the battery were standardized inventories; others were experimental in nature. The students took the tests at the beginning of the student-teaching semester, which occurred at the end of their senior year.

During the academic year 1954–55, a follow-up of the student teachers who were tested the year before and had subsequently received bachelors degrees was undertaken. Those students who were then teaching in Grades 3 to 6 in New York City public elementary schools in which at least one other member of the group was also teaching were encouraged to participate as subjects in an observational study. Of approximately 75 teachers who met these criteria, it was possible to conduct intensive observations in the classrooms of 49. In addition, several tests were administered to the pupils taught by these 49 teachers and to the teachers themselves.

This report will discuss three kinds of data:

1. Test scores of 49 student-teachers obtained during their senior year in college.

2. Classroom behavior records obtained through systematic observation approximately one year later in the classrooms of these 49 former student-teachers.

3. Scores on pupil-teacher rapport assigned to the 49 teachers on the basis of the reactions of their pupils to a paper-and-pencil attitudinal measure.

Test Scores

From the large group of tests taken by the student teachers, the authors selected the following tests which, on the basis of prior research and educational theory, could be expected to function as predictors of pupil-teacher rapport.

[1] This is one of a series of studies of teacher behavior currently being conducted by the Office of Research and Evaluation of the Division of Teacher Education of the Municipal Colleges of New York City. A longer version of the present paper may be had on request as long as the supply remains.

1. *The Minnesota Teacher Attitude Inventory (MTAI).* The MTAI was scored with two keys: the first, the published, empirically-derived key (**3**), and the second, an experimental key in which the items were scored on an a priori, rational basis.

2. *The California F Scale.* A 30-item version of the F scale was developed using the item analysis data in *The Authoritarian Personality* (**1**).

3. *The Draw-a-Teacher Technique (DaTt).* In the DaTt, a subject is given instructions to "draw a teacher with a class" (**10**). The drawings were scored by three scorers along three dimensions—Teacher Initiative, Psychological Distance, and Traditionalism in Classroom Organization (**9**). Interscorer agreement was estimated by analysis of variance procedures; the following intraclass correlations were obtained:

Teacher Initiative..................	.90
Psychological Distance..............	.93
Traditionalism in Classroom Organization	.84

4. *Sims SCI Occupational Rating Scale (SCI).* The SCI scale "is an instrument designed to reveal the level in our social structure—i.e. the social class—with which a person unconsciously identifies himself" (**11**, p. 1). A subject taking the SCI scale indicates whether he generally considers the people in each of 42 occupations (representative of varying levels of socioeconomic status) as belonging in the same, a higher, or a lower social class than he himself does.

5. *Strong Vocational Interest Blank (Index R).* Index R is a 95-item key developed by Mitzel (**8**) for the Strong Vocational Interest Blank for Women. This key is composed of those items which successfully discriminated high-rapport and low-rapport teachers (differentiated on the basis of principals' judgments and MTAI scores) and which survived cross-validation (based on extreme groups differentiated by the MTAI).

6. *Inventory IV—Satisfaction Score.* Inventory IV is an experimental inventory consisting of 32 multiple-choice items dealing with student-teaching experiences. It is scored to obtain a measure which, on the basis of the manifest content of the responses, appears to indicate the student-teacher's satisfaction with the student-teaching experience (**2**).

Measures of Classroom Behavior

A technique for observing and recording what occurs in a classroom, called the Observation Schedule and Record (OScAR), was developed to provide a means for objectively describing a variety of different classroom activities (**6**). The technique provides measures of the frequency of occurrence of specific classroom events, and requires few inferences on the part of the observer.

Each of the 49 teachers was observed by six different research workers. They observed each teacher for two one-half hour periods, adding up to a total of 588 observation periods. No two observers visited any given classroom at the same time.

From the basic behavioral data supplied by the OScAR technique, indices of teacher and pupil activities, types of pupil groupings, classroom climate, and expressive behavior of the teacher were derived. Of 14 dimensions developed, the following four were selected for study in this report because they seemed to be conceptually related to both the test measures and the criterion.

1. *Disorderly Pupil Behavior.* This dimension focuses on pupil behavior which reflects either hostility or disruptive activity (e.g., pupil ignores teacher's question, scuffles, etc.). It is a general index of the disorder present in a given classroom. Its reliability, determined by agreement among observers of the same class on different occasions, was estimated to be .89.

2. *Manifest Teacher Hostility.* This dimension provides an index of the overt, hostile, nonintegrative activity of the teacher. Verbal and nonverbal behaviors judged to reflect teacher hostility (e.g., sarcasm and scowling) were tallied and combined for this dimension. Its reliability was estimated to be .92.

3. *Pupil Leadership Activities.* This dimension provides an index of the amount of pupil leadership the teacher allows in classroom activity. It is based on activities in which a pupil addresses, or demonstrates to, the class. The reliability of this measure was estimated to be .72.

4. *Freedom of Movement.* This dimension offers an index of the freedom of movement exhibited by both pupil and teacher in the classroom. It reflects the teacher's apparent willingness to circulate among the pupils, and the ease with which a pupil can move about without requiring special permission. Its reliability was estimated to be .63.

Three other measures of the classroom were derived from global ratings of the classroom setting. The observers consulted the drawing scales developed and employed as part of the Draw-a-Teacher technique described earlier. After completing an observation period, the observer rated the class on each of the following dimensions: Teacher Initiative, Psychological Distance, and Traditionalism in Classroom Organization. The reliabilities of these ratings were estimated to be .72, .71, and .85, respectively.

Measure of Pupil-Teacher Rapport

In the present study, pupil-teacher rapport was defined as the generalized, conscious, subjective regard expressed by pupils for their teacher. In order to secure measures of the way in which the pupils perceived their teacher, an inventory, My Class, was constructed (**5**). This inventory consists of 47 scored items comprising four scales: Halo, Disorder, Supportive Behavior, and Traditionalism. The Halo scale is designed to indicate the extent to which the pupils have a general feeling of liking for the teacher, while the other three scales are intended to measure fairly specific teacher and pupil behaviors.

My Class was administered to all the pupils in the classes of the 49 teachers participating in the study. The items were read aloud to the pupils by a test administrator, while the teacher sat at the back of the room and filled out an inventory unrelated to the pupils' activity. The proportion of the class giving the keyed response to each item was used as the teacher's score on that item. A teacher's score on each scale was the sum of proportions for all of the items of that scale, appropriately weighted plus or minus.

The Halo scale consists of the following eight items scattered throughout the My Class inventory:

1. Do you ever feel like staying away from school?
2. Do you like to be in this class?
3. Do you have much fun in this class?
4. Do you learn a lot in this class?
5. Are you proud to be in this class?
6. Do you always do your best in this class?
7. Do most of the pupils like the teacher?
8. Does the teacher help you enough?

The reliability of this scale, estimated by analysis of variance procedures, was .89.

Results

For each teacher in this study, there were 17 measures:[2] nine test scores, seven classroom observation measures, and one measure of pupil-teacher rapport based on pupil reactions. The primary analysis of these data consisted of correlating each of these measures with the other 16. The resulting correlations are contained in Table 1. From an examination of Table 1 it is clear that:

1. None of the tests correlates significantly with the measure of pupil-teacher rapport.
2. None of the 63 correlations between the test variables and the classroom behavior variables is significant except that between the Teacher Initiative score on the DaTt and the OScAR dimension, Freedom of Movement.
3. The only classroom behavior variable that correlates significantly with the Halo

[2] This is not strictly speaking the case, since complete test data were not available for every one of the 49 subjects. See footnote *a* on Table 1.

score of My Class is the OScAR dimension Manifest Teacher Hostility. The correlation is in the "expected" direction, i.e., the pupils' liking for their teacher as indicated on My Class decreases with the amount of manifest teacher hostility recorded by observers. Evidently the more hostility a teacher displays in the classroom, the less esteemed she is by her pupils.

A multiple regression analysis was employed using the pupil-teacher rapport criterion with all of the test variables except the MTAI-Rational Key score as independent variables. When weighted optimally with the partial regression coefficients, the eight test scores correlated .496 with Halo. This multiple correlation coefficient is not significant.

Discussion

The major finding of this study is the failure of the tests, singly or in combination with one another, to predict subsequent pupil-teacher rapport as measured by the Halo scale. Each of the tests was selected for study because theory or past research, and sometimes both, encouraged its use as a potential predictor. The fact that none of the tests adequately functioned to predict pupil-teacher rapport is therefore of particular interest.

One of the distinguishing features of this study is that the tests were administered to a group of college seniors who had not yet served as teachers. Since the tests and criterion were well separated in time, the study deals with the predictive, rather than concurrent, validities of the tests employed. In general, the results offer no evidence of the predictive validity of any of the tests for the particular criterion measure studied. The tests not only failed to predict rapport, they did not correlate with the objective measures of behavior in the classroom. Of the 63 correlations between test variables and classroom behavior variables, only the relationship between the Teacher Initiative dimension of the DaTt and the Freedom of Movement dimension of OScAR proved significant.

The fact that a test has concurrent validity is often incorrectly used to support a recommendation for its use as a predictive measure. Thus, the MTAI has been shown to correlate with various independent measures of pupil-teacher rapport (**3**), including measures based on pupil responses to questions such as were used in My Class. The well-established concurrent validity of the MTAI does not, however, demonstrate that the test is of predictive value, and the recommendation contained in the test manual that the inventory be used as a predictor is accordingly without empirical support. The evidence of the present investigation, which is the only published research of which the writers are aware involving the correlation of the MTAI and a subsequent measure of pupil-teacher rapport, would argue strongly against its use as a predictive instrument.

It may be important to note that the pupil-teacher rapport criterion used in this study was not so uniquely or completely determined by the personalities of the pupils who responded to My Class as to be unrelated to measurable, behavioral variables in the teacher. As Table 1 indicates, one of the measures of the teachers' classroom behavior, Manifest Teacher Hostility, correlates significantly with the criterion. Moreover, in a previously reported investigation, Medley and Williams (**7**) found that the Halo scores of the 49 teachers in this study correlated significantly ($r = +.34$) with their scores on a concurrent test measure of hostility.[3] Since the criterion used in

[3] It is of interest to note that the Hostility scale was built by selecting 50 items from among those on the Minnesota Multiphasic Personality Inventory that were found to discriminate significantly between teachers

this study is correlated with a concurrent measure of the teachers' classroom behavior and test behavior, the failure of the predictive instruments cannot be attributed to inherent unpredictability of the criterion.

In the past, demonstrations of the concurrent validity of tests have, too often, been uncritically accepted as evidence of their predictive value. The study reported here, however, adds support to the growing view that the predictive value of tests can only be established through predictive studies.

Summary

A large group of student teachers were given a number of personality and attitude tests during their senior year in college. Observations were conducted approximately one year later in the rooms of 49 of these subjects who were employed as elementary school teachers. A measure of pupil-teacher rapport based on pupil responses to questions about their teacher and their class was also obtained.

In general, none of the test measures correlated significantly with pupil-teacher rapport as measured. Only one of the 63 correlations between the test measures and classroom behavior measures proved significant. Manifest Teacher Hostility, a measure based on classroom observation of the teacher correlated significantly with rapport.

The implications of these results for the prediction of pupil-teacher rapport were discussed.

scoring high and low on the MTAI (**4**). In view of the manner in which the Hostility scale was developed, it is difficult to determine why it should correlate significantly with the Halo scale while the MTAI does not. Only when the temporal relations of the Hostility scale and the MTAI to the criterion are fully appreciated does the difference in the validity coefficients become understandable.

References

1. Adorno, T. W., Frenkel-Brunswik, E., Levinson, D. J., & Sanford, R. N. *The authoritarian personality.* New York: Harper, 1950.
2. Aikman, L. P., & Ostreicher, L. M. *Development of an inventory for measuring satisfaction with student teaching.* Research Series No. 22, Division of Teacher Educ., Board of Higher Educ., City of New York, July 1954.
3. Cook, W. W., Leeds, C. H., & Callis, R. *Minnesota Teacher Attitude Inventory-Manual.* New York: Psychological Corp., 1951.
4. Cook, W. W., & Medley, D. M. Proposed hostility and pharisaic virtue scales for the MMPI. *J. appl. Psychol.,* 1954, **38,** 414–418.
5. Medley, D. M., & Klein, Alix A. *Studies of teacher behavior: inferring classroom behavior from pupil responses.* Research Series No. 30, Division of Teacher Educ., Board of Higher Educ., City of New York, February 1956.
6. Medley, D. M. & Mitzel, H. E. A technique for measuring classroom behavior. *J. educ. Psychol.,* 1958, **49,** 86–92.
7. Medley, D. M., & Williams, Ida F. *Predicting teacher effectiveness with the Minnesota Multiphasic Personality Inventory.* Research Series No. 34, Division of Teacher Educ., Board of Higher Educ., City of New York, February 1957.
8. Mitzel, H. E. Interest factors predictive of teacher's rapport with pupils. Unpublished doctoral dissertation, Univer. Minnesota, 1952.
9. Mitzel, H. E., Ostreicher, L. M., & Reiter, S. R. *Development of attitudinal dimensions from teachers' drawings.* Research Series No. 24, Division of Teacher Educ., Board of Higher Educ., City of New York, October 1954.
10. Rabinowitz, W., & Travers, R. M. W. A drawing technique for studying certain outcomes of teacher education. *J. educ. Psychol.,* 1955, **46,** 257–273.
11. Sims, V. M. *Sims SCI Occupational Rating Scale.* New York: World Book, 1952.

Received December 11, 1957.

Journal of Educational Psychology
Vol. 49, No. 2, 1958

A RECALCULATION OF FOUR ADULT READABILITY FORMULAS

R. D. POWERS, W. A. SUMNER, AND B. E. KEARL

Department of Agricultural Journalism, University of Wisconsin

When the so-called readability formulas are used only as rough estimating devices for the encouragement of popular writing, statistical precision is not vitally important. But if they are to be considered research tools in studies of comprehension or learning, it becomes very important to build into them as much precision as possible.

Current readability formulas offer at least two opportunities for reexamination for the sake of greater precision. First, many are based on reading comprehension tests published in 1926 and drawn from empirical testing of school pupils prior to that date. Thus they may not adequately reflect changes in either the language or the population of the present decade. Second, the "ratings" produced by present tests are not accompanied by a standard error figure, and hence tell nothing about significance of estimates and differences.

The revision of the set of graded test passages used in building two widely used readability formulas—the Flesch Reading Ease Formula (**3**) and the Dale-Chall readability formula (**1**)—has offered an opportunity for revision of the formulas and also for further comparative evaluation of these two indexes of comprehension difficulty. The two formulas were originally calculated, following Lorge (**6**), by making measurements of sentence length and vocabulary difficulty in the 1926 edition of the McCall-Crabbs Graded Test Lessons in Reading (**7**). Both formulas make use of the same sentence length measure—average number of words per sentence. For a vocabulary measurement, Flesch uses the number of syllables per 100 words, while Dale and Chall count the number of words that do not appear on a list of 3,000 words which had proved "familiar" for youngsters tested in the fourth grade of public schools.

Results with the two formulas are not directly comparable for several reasons. Scores are given in different terms—Flesch results on a scale of 100 (easy) to 0 (difficult) and Dale-Chall results on a scale of about 3 (easy) to 14 (difficult). Flesch's formula was calculated with grouped data, while Dale and Chall computed theirs with ungrouped data. In addition, Flesch made an adjustment in one of the formula terms after computation, while Dale and Chall did not.[1]

More recent arrivals on the readability scene are the Farr-Jenkins-Paterson simplification of the Flesch Reading Ease formula (**2**) and the Gunning Fog Index (**4**). The former uses a count of percentage of monosyllables instead of the Flesch syllable count, while the latter uses a count of polysyllables (words of more than two syllables). Both formulas take sentence length into account. Both are viewed by many as simplifications of the Flesch formula.

The McCall-Crabbs tests were revised considerably in 1950 (**8**). There is evidence that the questions and passages of the 1926 edition were changed considerably in the 1950 edition. At least 60 of the tests in the 1950 edition are different in subject from those in the earlier edition.

[1] This adjustment concerned the criterion value used in developing the regression equation. Both Flesch and Dale-Chall used as a criterion the average school grade of pupils answering correctly 50% of the questions accompanying the reading passages. But Flesch adjusted the formula to predict the grade of the pupil who could answer 75% of the questions correctly. This changed the regression formula constant.

TABLE 1

AVERAGES AND STANDARD DEVIATIONS OF MEASUREMENTS IN TWO EDITIONS OF THE McCALL-CRABBS STANDARD TEST LESSONS IN READING

	1950 Edition	1926 Edition	
		Flesch	Dale-Chall
Mean average grade of pupils answering 50% correctly (criterion)	4.9862 ($s = 1.1068$)	5.4973 ($s = 1.3877$)	5.7492 ($s = 1.6565$)
Average number of words per sentence	15.3986 ($s = 3.8373$)	16.5213 ($s = 5.5509$)	16.8037 ($s = 5.3818$)
Average syllables per hundred words	131.6131 ($s = 11.830$)	134.2208 ($s = 13.6845$)	— —
Average percentage of words not on Dale list	6.9413 ($s = 5.8200$)	— —	8.1011 ($s = 6.3056$)
Average percentage monosyllables	75.1148 ($s = 6.8083$)	— —	— —
Average percentage polysyllables	5.7603 ($s = 4.5885$)	— —	— —

Those are the passages dealing with World War II, atomic energy, modern aviation, and similar recent developments. Table 1 shows how the two editions differed in averages and standard deviations of the various measurements.

PURPOSE OF REVISION

It was felt that recalculation of these four formulas with the 1950 tests as a criterion would accomplish two main purposes: (*a*) modernize the formulas by taking advantage of the more recently administered tests which should reflect some of the changes in pupil reading abilities between 1926 and 1950, and (*b*) establish formulas which are derived from identical materials, measured by identical rules of measurement on the common factor, calculated by identical mathematical operations, and reported without adjustment. The latter goal seems desirable because it will make further comparative studies easier to perform and interpret (i.e., no manipulations of the recalculated formulas will be needed in future research toward modernization and validation). It would also allow averaging of several formula results for any sample of writing, thus perhaps giving more accurate scores where extreme accuracy is needed.

METHODS

The following measurements were made in the 383 prose passages of the 1950 edition of the McCall-Crabbs tests:

1. Average grade score of pupils answering half the test questions correctly.
2. Average number of words per sentence in each passage.
3. Number of syllables per 100 words in each passage.
4. Percentage of words in each passage not appearing on Dale's list of 3,000 "easy" words.
5. Percentage of monosyllables in each passage.
6. Percentage of polysyllables in each passage.

Regression formulas were computed with these measurements,[2] and the results of the formulas were compared by applying them to 113 samples of writing from various publications to determine the practical significance of differences in formula results. The recalculated Flesch and Dale-Chall formulas were also compared with each other and with results from the original formulas[3] in a sample of 40 of the McCall-Crabbs passages. Such comparisons with the other recalculated for-

[2] Calculation facilities used were in the Wisconsin Numerical Research Laboratory, supported by a National Science Foundation grant and funds from the Wisconsin Alumni Research Foundation allocated by the University of Wisconsin Graduate School Research Committee.

[3] The original formulas were as follows: Flesch: 206.84 − (1.015)(sent. length) − (.846)(syllables per 100 words). Dale-Chall: 3.6365 + (.0496)(sent. length) + (.1597) (% non-Dale words). Gunning: .4 (sent. length + % poylsyllables). F-J-P: −31.517 − (1.015)(sent. length) + (1.599)(% mono-

mulas were not possible because adjustments of the original formulas were not possible or because different rules for word counting were used in the original formula and the recalculation.

Results

The calculations yielded the following recalculated formulas:

- Flesch: $-2.2029 + (.0778)$ (sentence length) $+ (.0455)$ (syllables per 100 words)
- Dale-Chall: $3.2672 + (.0596)$ (sentence length) $+ (.1155)$ (% non-Dale words)
- Farr-Jenkins-Paterson: $8.4335 + (.0923)$ (sentence length) $- (.0648)$ (% monosyllables)
- Gunning Fog Index: $3.0680 + (.0877)$ (sentence length) $+ (.0984)$ (% polysyllables)

The coefficients of multiple determination ($\overline{R}^2$)—which indicates the amount of variation in difficulty among the tests which is accounted for by the two style variables in the formula—are .4034 for the recalculated Flesch formula, .5092 for the recalculated Dale-Chall formula, .3407 for the Farr-Jenkins-Paterson recalculation, and .3440 for the Gunning recalculation. These statistics, which are corrected for degrees of freedom, show that the recalculated Flesch formula statistically "explains" some 40% of the variation in difficulty of the McCall-Crabbs tests. The Dale-Chall formula explains almost 51%, and is thus the much more powerful tool for predicting reading difficulty. The Farr-Jenkins-Paterson and Gunning formulas as recalculated are about equal in predictive power—both considerably weaker than the other formulas.

The error terms for the formulas are .85 school grades for the Flesch formula, .77 grades for the Dale-Chall formula, and .90 grades for the others. Converting the predicted value for each formula into a grade level figure and following the standard practice of taking a range of plus or minus two standard errors as the probable area in which the "true" value lies, the error range would be 1.71 grades for the Flesch formula, 1.55 grades for the Dale-Chall formula, and 1.80 grades for both the others. Thus the Dale-Chall formula came through the recalculations as slightly more precise than the others.

Table 2 presents comparisons of various statistics for the Flesch formula and Dale-Chall formula in their recalculated and original forms and for the recalculated Farr-Jenkins-Paterson simplification and Gunning Index.

To assess the practical significance of the revision, the original and recalculated forms of the Flesch and Dale-Chall formulas were applied to 47 sample passages from a variety of sources. The recalculated Dale-Chall formula consistently gave lower scores than the original; the average discrepancy (average absolute deviation) between the two was .94 grades. The average discrepancy between the original and recalculated Flesch formulas was .85 grades, with the recalculated formula giving a lower score about two-thirds of the time.

All four recalculated formulas were compared in a sample of 113 passages from 15 magazines. The results are given in Table 3. The writers feel that two observations from the table are worthy of mention here. First, the average discrepancy of results using the recalculated Flesch and Dale-Chall formulas was .54 school grades.

syllables). The Flesch formula used for comparison with the new one had to be adjusted back to predict at the 50% level of the criterion and the scale reversed (i.e., changed back to the form which was presumably yielded directly by Flesch's computations before he made the various adjustments. This unscaled, reversed formula, as nearly as we can determine, is $-7.5695 + (.1015)$(sent. length) $+ (.0864)$(syllables per 100 words). It was not possible to put the Farr-Jenkins-Paterson simplification on such a basis.

TABLE 2

Regression Statistics for the Recalculated and Original Formulas

Statistic[a]	Flesch Formula		Dale-Chall Formula		F-J-P Formula	Gunning Index
	Recalculated	Original	Recalculated	Original	Recalculated	Recalculated
r^2_{12}	.2019	.2695	.2019	.2191	.2019	.2019
r^2_{13}	.3436	.4420	.4759	.4670	.2526	.2665
r^2_{23}	.1363	.2157	.1599	.2607	.1055	.1265
$r^2_{12.3}$	.0987	.1743+	.0450	.1331+	.1146	.1013
$r^2_{13.2}$	.3117	.2202+	.3883	.1936+	−.6293	.1984
$b_{12.3}$	.0778	.1015	.0596	.0496	.0923	.0877
$b_{13.2}$	.0455	.0846	.1155	.1579	−.0648	.0984
$\beta_{12.3}$	.2697	.2639	.2065	.1611	.3199	.3042
$\beta_{13.2}$	.4865	.5422	.6073	.6011	.3986	.4081
$a_{1.23}$	−2.2029	−7.5695	3.2790	3.6365	8.4335	3.0680
$R^2_{1.23}$	.4034	.4966	.5092	.4900	.3407	.3440

Note.—+ Values computed from those given by Flesch or Dale and Chall by the relationship:

$$r_{ij.k} = \frac{r_{ij} - r_{ik}r_{jk}}{\sqrt{(1 - r_{ik}^2)(1 - r_{jk}^2)}}$$

[a] Subscripts refer to (1) the criterion, (2) average sentence length. Subscript (3) refers to a different variable for each formula: syllable per 100 words for Flesch formula, percentage words not on Dale List for the Dale-Chall formula, percentage monosyllables for the Farr-Jenkins-Paterson formula, and percentage polysyllables for the Gunning index.

In the comparison of the original Dale-Chall and Flesch formulas above, the average discrepancy was .87 grades. All four recalculated formulas agreed much more closely with one another than the original Dale-Chall and Flesch formulas did. This would seem to be a point in favor of the recalculations.

TABLE 3

Comparisons Between Results with Recalculated Formulas Applied to 113 100-word Samples of Prose

Comparison	Positive deviations		Negative deviations		Average absolute deviation
	size	(number)	size	(number)	
Dale-Chall and Flesch	.51	(58)	.57	(55)	.54
Flesch and Gunning	.54	(82)	.15	(31)	.44
Dale-Chall and Gunning	.65	(82)	.31	(29)	.56
Flesch and F-J-P	.56	(96)	.13	(16)	.50
Dale-Chall and F-J-P	.73	(93)	.35	(20)	.66
Gunning and F-J-P	.36	(70)	.57	(42)	.54

The recalculated Gunning Index gave results that were in slightly higher agreement with the results of the recalculated Flesch formula than were the results with the recalculated Farr-Jenkins-Paterson simplification. The average absolute deviation between the recalculated Flesch formula and Gunning Index was .44 grades, with 73% of the predictions lower than those of the Flesch formula. The average absolute deviation between the recalculated Flesch formula and the recalculated Farr-Jenkins-Paterson simplification was .50 grades, with 85% of the results with the simplification being lower than results with the Flesch formula. Thus the two simplifications gave slightly lower scores than the recalculated Flesch formula. Scores with the recalculated Gunning Index were slightly closer to the Flesch results, and there were more instances of predictions which were higher than the Flesch predictions than was true of the Farr-Jenkins-Paterson formula as recalculated.

TABLE 4

NORMS OF RECALCULATED FORMULA SCORES FOR MATERIAL OF VARIOUS TYPES

Material	Style		Flesch	Dale-Chall	Gunning	Farr-Jenkins-Paterson
Scientific:						
Phytopathology, Soil Science, Journal of Nutrition, Science, American Journal of Veterinary Research	Difficult	Means...	8.00	8.50	7.70	7.20
		Ranges..	7.10–9.50	7.10–10.70	6.70–8.90	6.30–7.40
Academic:						
Yale Review, Harvard Educational Review, Annuals of the American Academy of Political and Social Sciences.	Difficult	Means...	7.90	8.40	7.60	7.00
		Ranges..	7.00–8.70	7.50–8.60	7.10–8.60	6.50–8.00
Quality:						
Harper's, Atlantic Monthly	Fairly Difficult	Means...	6.80	6.70	6.40	6.50
		Ranges..	6.30–7.60	5.50–8.60	5.40–8.50	5.40–6.80
Standard:						
Reader's Digest	Average	Means...	6.00	6.10	5.70	5.90
		Ranges..	5.00–7.20	4.80–7.00	5.00–6.70	4.90–6.70
Slick Fiction:						
Colliers, Ladies Home Journal, Good Housekeeping	Fairly Easy	Means...	4.90	4.90	4.90	4.90
		Ranges..	4.30–6.20	4.50–6.70	4.40–6.20	4.40–6.50
Pulp Fiction:						
True Confessions	Easy	Means...	4.30	4.20	4.30	4.20
		Ranges..	3.70–4.40	3.70–4.50	3.80–4.80	3.70–4.50

The application of the formulas to the 113 passages also provides some "norms" for interpreting scores which they yield. The passages came from various types of publications, presumably representing generally different levels of reading difficulty as noted in Table 4.

Use of such a scale is admittedly a rough manner of interpreting readability scores, and the scale in Table 4 was not formed in the most exact manner; although sampling was at random within issues, the issues were not randomly chosen and calculations were rounded. However, this general approach seems more desirable than using the theoretical formula result (grade level) or making adjustments in the theoretical result without benefit of extended testing. Further details on background, method, and results of this work are available on microfilm (**9**).

CONCLUSIONS AND DISCUSSION

To recommend use of the four recalculated formulas in preference to the original ones is a rather drastic step, in view of the wide use the original formulas have enjoyed. However, such a recommendation is made here for the reasons we set forth in the paragraphs on the purpose of the recalculation.

The formula coefficients derived in the recalculations on the 1950 McCall-Crabbs tests have the same statistical validity as those calculated on the 1926 edition of the tests. They are statistically preferable to those formed by rougher, short-cut procedures.

Reservations in making a recommendation to use the recalculated Dale-Chall and Flesch formulas stem from two ¯basic sources: (*a*) Readability formulas are such rough estimates at best that to say one result is better than another is statistically hazardous—especially when the nature of the material on which the formulas are to be used differs from that of the material used in computing the formula. (*b*) In the revision of the McCall-Crabbs criterion tests, passages of higher difficulty were omitted. The style measurements of these passages and the educational level of pupils taking these more difficult passages might have been of a type which more nearly approaches the type of writing and audience for which the formulas are normally used. In other words, restriction of the range of difficulty in the 1950 tests may have made this edition less suitable than the 1926 tests for building readability formulas. But to the extent this argument is sound, all linear formulas suffer equally from the curvilinearity it implies.

It is further recommended that the Dale-Chall formula be used whenever possible in the absence of specific reasons for preferring the Flesch formula or one of its simplifications. The Dale-Chall formula was best in terms of small error and high prediction power. This parallels an earlier judgment by Klare (**5**) that the original Dale-Chall formula was better than the original Flesch formula by a slight margin.

The statements here as to error and prediction power of the formulas apply only to prediction and precision in regard to the criterion passages. They do not unequivocally hold true for the formulas as they are normally used—for estimating difficulty of adult reading materials. It is possible that a formula with low precision or predictive power in this research could be fully as precise as the others for predicting adult reading difficulty. But there is no direct evidence that this would be so, and the only recourse at present seems to be to give the Dale-Chall formula the highest place on the basis of its prediction power and small error computed on the criterion.

Some formula-users—particularly those who use formulas only occasionally—are understandingly reluctant about referring to a word list, which is required by the Dale-Chall formula. Of popular formulas without word lists, the Flesch formula is statistically best.

Those who use either simplification of the Flesch formula should recognize that they are sacrificing precision and accuracy by doing so. But it seems evident that for estimates of readability which need to be performed rapidly and where precision is not extremely important, either simplification will do the job.

There are two ways of looking at precision in a readability formula. One way is to admit that formulas are rough estimates at best, and that a loss of a little precision is not important. The other is to argue that since the formulas give only rough estimates, it is important to keep whatever precision and prediction power exists.

The choice of viewpoint seems to hinge on the use to be made of formula results. A news writer or editor who uses a formula "to see how we are doing" could probably regard all four formulas as equal for his purpose and use whichever formula he found easiest to apply. If readability scores are part of a research design, however, the social scientist will want to choose the most powerful and precise formula even though it entails more difficulties in application.

REFERENCES

1. Dale, E., & Chall, J. S. A formula for predicting readability. *Education Res. Bull.* Ohio State Univer., 1948, **27,** 11–20, 37–54.
2. Farr, J. N., Jenkins, J. J., & Paterson, D. G. Simplification of the Flesch Reading Ease Formula. *J. appl. Psychol.,* 1951, **35,** 333–337.

3. Flesch, R. A new readability yardstick. *J. appl. Psychol.*, 1948, **32**, 221–233.
4. Gunning, R. *The technique of clear writing*. New York: McGraw-Hill, 1952.
5. Klare, G. R. Measures of the readability of written communication: An evaluation. *J. educ. Psychol.*, 1952, **43**, 385–399.
6. Lorge, I. Predicting reading difficulties in selections for children. *Elem. English Rev.*, 1939, **16**, 229–233.
7. McCall, W. A., & Crabbs, L. M. *Standard test lessons in reading*. New York: Columbia Univer. Teachers College, 1926.
8. McCall, W. A., & Crabbs, L. M. *Standard test lessons in reading*. New York: Columbia Univer. Teachers College, 1950.
9. Powers, R. D. A recalculation and partial validation of four adult reading formulas. Unpublished doctoral dissertation, Univer. of Wisconsin, 1957. (Also available from University Microfilms, Ann Arbor, Michigan.)

Received December 30, 1957.

Journal of Educational Psychology
Vol. 49, No. 2, 1958

NOTE ON THE ORGANISMIC AGE CONCEPT

PAUL BLOMMERS AND J. B. STROUD
State University of Iowa

In the March 1955 issue of this journal the authors, together with Knief, presented data showing that the use of height age, weight age, and dental age contributed practically nothing to a least squares estimate of either reading or arithmetic achievement when combined with mental age (**1**). That is to say, mental age alone provided about as accurate an estimate of achievement in these areas as a least squares combination of mental age and these three other age scores.

It is well-known that for the model assumed (usually a first degree polynomial) a least squares combination of scores provides composite scores for the individuals at hand which bear a maximum degree of relationship to the criterion. Hence, multiple correlations between achievement and the component variables that enter into organismic age (OA) cannot be lower than the simple correlations between achievement and mental age (MA) alone. In such multiple correlation analyses the component age scores are, of course, automatically ideally weighted before being combined. In the formation of the OA score, on the other hand, the component age scores are given equal weights[1] since the OA score is the simple unweighted average of the component age scores. Because of the nature of the relationships among these age scores it follows that the correlation between educational achievement and OA must necessarily be considerably lower than that between achievement and MA alone. The purpose of this note is to demonstrate this fact analytically. We shall also use data reported in our previous article to illustrate the extent of this attenuating effect of anatomical and physiological age scores when used in combination with MA to predict school achievement.

In the following discussion and in keeping with our previous article, we shall consider as estimators only mental, height, weight, and dental age scores. We shall designate these as M, H, W, and D, respectively. Reading and arithmetic scores will be used as measures of school achievement and will be designated R and A, respectively. Since the correlation between the sum of the four age scores and either R or A is identical with the correlation between the mean of these four scores and R or A, we shall discuss the efficacy of OA as a predictor of R or A where OA = O = M + H + W + D. The symbol Cov RO will be used to refer to the covariance between R and O scores while the symbol Var O will be used to refer to the variance of the O scores.

Consider reading (R) as the criterion. Then

$$\begin{aligned}\text{Cov RO} &= \text{Cov R(M + H + W + D)}\\ &= \text{Cov RM} + \text{Cov RH}\\ &\quad + \text{Cov RW} + \text{Cov RD}.\end{aligned} \qquad [1]$$

But the various anatomical and physiological scores tend to bear a very low degree of relationship to reading achievement so that Cov RM is large in relation to Cov RH + Cov RW + Cov RD. That is, the addition of H, W, and D to M does not supplement Cov RM to any marked extent, so that Cov RM accounts for most of Cov RO.

[1] By a weight we mean the constant by which the score for a trait is multiplied before it is combined with other similarly weighted scores to form a composite.

Next note that

$$\begin{aligned}\text{Var O} &= \text{Var(M + H + W + D)}\\ &= \text{Var M + Var H + Var W}\\ &\quad + \text{Var D} + 2\ \text{Cov MH}\\ &\quad + 2\ \text{Cov MW} + 2\ \text{Cov MD}\\ &\quad + 2\ \text{Cov HW} + 2\ \text{Cov HD}\\ &\quad + 2\ \text{Cov WD}.\end{aligned} \qquad [2]$$

It is clear from [2] that even if the covariances involving M with H, W, and D are small, the variance of M is a relatively small portion of the variance of O. Now the correlation between R and M is given by

$$\frac{\text{Cov RM}}{\sqrt{(\text{Var R})(\text{Var M})}}, \qquad [3]$$

while the correlation between R and O is given by

$$\frac{\text{Cov RO}}{\sqrt{(\text{Var R})(\text{Var O})}}. \qquad [4]$$

As we have indicated, the numerator of [3] differs little from that of [4], while the denominator of [4] is much greater than that of [3] due to the fact that Var O must necessarily be much greater than Var M. Hence it is a simple mathematical fact that the correlation between R and O must be less than that between R and M. It, of course, follows in general that mental age alone is a much more useful predictor of school achievement than is mental age in equally weighted combination with various anatomical and physiological age scores, that is, OA. Moreover, the more physiological and anatomical age scores used in determining OA, the greater the attenuation of the correlation between OA and a school achievement criterion.

To show how marked this attenuation actually becomes, Formulas [1], [2], and [4] were applied to the products matrix used in the least squares analysis reported in our previous paper. In this paper the correlation between R and M was reported as .645. When H, W, and D are added to M to form O, the correlation of R with O is only .24. With arithmetic achievement as the criterion, the correlation between A and M previously reported was .551. In this case when H, W, and D are added to M to form O, the correlation of A with O is .21.

In brief, there are neither theoretical nor empirical bases for believing that organismic age predicts school achievement. This is not to say that OA may not be useful in predicting other types of behavior. However, evidence of such usefulness is not as yet generally available.

REFERENCES

1. Blommers, P.; Knief, Lotus, & Stroud, J. B., The organismic age concept, *J. educ. Psychol.*, 1955, **46**, 142–150.

Received January 17, 1958.

THE JOURNAL OF

EDUCATIONAL PSYCHOLOGY

Volume 49 August, 1958 Number 4

TEACHER COMMENTS AND STUDENT PERFORMANCE: A SEVENTY-FOUR CLASSROOM EXPERIMENT IN SCHOOL MOTIVATION[1]

ELLIS BATTEN PAGE

University of California, Los Angeles[2]

Each year teachers spend millions of hours marking and writing comments upon papers being returned to students, apparently in the belief that their words will produce some result, in student performance, superior to that obtained without such words. Yet on this point solid experimental evidence, obtained under genuine classroom conditions, has been conspicuously absent. Consequently each teacher is free to do as he likes; one will comment copiously, another not at all. And each believes himself to be right.

The present experiment investigated the questions: 1. Do teacher comments cause a significant improvement in student performance? 2. If comments have an effect, which comments have more than others, and what are the conditions, in students and class, conducive to such effect? The questions are obviously important for secondary education, educational psychology, learning theory, and the pressing concern of how a teacher can most effectively spend his time.

Previous Related Work

Previous investigations of "praise" and "blame," however fruitful for the general psychologist, have for the educator been encumbered by certain weaknesses: Treatments have been administered by persons who were extraneous to the normal class situation. Tests have been of a contrived nature in order to keep students (unrealistically) ignorant of the true comparative quality of their work. Comments of praise or blame have been administered on a random basis, unlike the classroom where their administration is not at all random. Subjects have often lacked any independent measures of their performance, unlike students in the classroom. Areas of training have often been those considered so fresh that the students would have little previous history of related success or failure, an assumption impossible to make in the classroom. There have furthermore been certain statistical errors: tests of significance have been conducted as if students were totally independent of one another, when in truth they were interacting members of a small number

[1] Portions of this paper were read at the National Research Conference of the American Educational Research Association at San Francisco, March 8, 1958. This research depended upon cooperation from many persons. Space limitations prevent the listing of their names. The writer is especially indebted to the teachers who freely donated time and energy after having been randomly [illegible] ut, their participation the study obviously would have been impossible.

[2] Where this study was conceived as part of a doctoral dissertation. The study was conducted in the San Diego City and County Schools while the writer was with San Diego Junior College. He is presently Coordinator of Guidance, Eastern Michigan College.

of groups with, very probably, some group effects upon the experimental outcome.

For the educator such experimental deviations from ordinary classroom conditions have some grave implications, explored elsewhere by the present writer (**5**). Where the conditions are highly contrived, no matter how tight the *controls*, efforts to apply the findings to the ordinary teacher-pupil relationship are at best rather tenuous. This study was therefore intended to fill both a psychological and methodological lack by *leaving the total classroom procedures exactly what they would have been without the experiment*, except for the written comments themselves.

METHOD

Assigning the subjects. Seventy-four teachers, randomly selected from among the secondary teachers of three districts, followed detailed printed instructions in conducting the experiment. By random procedures each teacher chose one class to be subject from among his available classes.[3] As one might expect, these classes represented about equally all secondary grades from seventh through twelfth, and most of the secondary subject-matter fields. They contained 2,139 individual students.

First the teacher administered whatever objective test would ordinarily come next in his course of study; it might be arithmetic, spelling, civics, or whatever. He collected and marked these tests in his usual way, so that each paper exhibited a numerical score and, on the basis of the score, the appropriate letter grade A, B, C, D, or F, each teacher following his usual policy of grade distribution. Next, the teacher placed the papers in numerical rank order, with the best paper on top. He rolled a specially marked die to assign the top paper to the *No Comment, Free Comment*, or *Specified Comment* group. He rolled again, assigning the second-best paper to one of the two remaining groups. He automatically assigned the third-best paper to the one treatment group remaining. He then repeated the process of rolling and assigning with the next three papers in the class, and so on until all students were assigned.

Administering treatments. The teacher returned *all* test papers with the numerical score and letter grade, as earned. No Comment students received nothing else. Free Comment students received, in addition, whatever comment the teacher might feel it desirable to make. Teachers were instructed: "Write anything that occurs to you in the circumstances. There is not any 'right' or 'wrong' comment for this study. A comment is 'right' for the study if it conforms with your own feelings and practices." Specified Comment students, regardless of teacher or student differences, all received comments designated in advance for each letter grade, as follows:

A: Excellent! Keep it up.
B: Good work. Keep at it.
C: Perhaps try to do still better?
D: Let's bring this up.
F: Let's raise this grade!

Teachers were instructed to administer the comments "rapidly and automatically, trying not even to notice who the students are." This instruction was to prevent any extra attention to the Specified Comment students, in class or out, which might confound the experimental results. After the comments were written on each paper and recorded on the special sheet for the experimenter, the test papers were returned to the students in the teacher's customary way.

It is interesting to note that the student subjects were totally naive. In other psychological experiments, while often not aware of precisely what is being tested,

[3] Certain classes, like certain teachers, would be ineligible for a priori reasons: giving no objective tests, etc.

subjects are almost always sure that something unusual is underway. In 69 of the present classes there was no discussion by teacher or student of the comments being returned. In the remaining five the teachers gave ordinary brief instructions to "notice comments" and "profit by them," or similar remarks. In none of the classes were students reported to seem aware or suspicious that they were experimental subjects.

Criterion. Comment effects were judged by the scores achieved on the very next objective test given in the class, regardless of the nature of that test. Since the 74 testing instruments would naturally differ sharply from each other in subject matter, length, difficulty, and every other testing variable, they obviously presented some rather unusual problems. When the tests were regarded primarily as *ranking* instruments, however, some of the difficulties disappeared.

A class with 30 useful students, for example, formed just 10 levels on the basis of scores from the first test. Each level consisted of three students, with each student receiving a different treatment: No Comment, Free Comment, or Specified Comment. Students then achieved new scores on the second (criterion) test, as might be illustrated in Table 1, Part A. On the basis of such scores, they were assigned rankings within levels, as illustrated in Table 1, Part B.

TABLE 1

Illustration of Ranked Data

Level	Part A (Raw scores on second test)			Part B (Ranks-within-levels on second test)		
	N	F	S	N	F	S
1	33	31	34	2	1	3
2	30	25	32	2	1	3
3	29	33	23	2	3	1
...	...	...	...	...	...	...
...	...	...	...	...	...	...
...	...	...	...	...	...	...
10	14	25	21	1	3	2
Sum:				19	21	20

Note.—N is No Comment; F is Free Comment; S is Specified Comment.

If the comments had no effects, the sums of ranks of Part B would not differ except by chance, and the two-way analysis of variance by ranks would be used to determine whether such differences exceeded chance.[4] Then the *sums* of ranks themselves could be ranked. (In Part B the rankings would be 1, 3, and 2 for Groups N, F, and S; the highest score is ranked 3 throughout the study.) And a new test, of the same type, could be made of all such rankings from the 74 experimental classrooms. Such a test was for the present design the better alternative, since it allowed for the likelihood of "Type G errors" (**3**, pp. 9–10) in the experimental outcome. Still a third way remained to use these rankings. The summation of each column could be divided by the number of levels in the class, and the result was

[4] The present study employed a new formula,

$$\chi_r^2 = \frac{6\Sigma(0 - E)^2}{\Sigma 0}$$

which represents a simplification of Friedman's twenty-year-old notation (**2**). The new form is the classic chi square,

$$\Sigma \frac{(0 - E)^2}{E}$$

multiplied by $6/k$ where k is simply the number of ranks! This conversion was discovered in connection with the present study by a collaboration of the writer with Alan Waterman and David Wiley. Proof that it is identical with the earlier and more cumbersome variation,

$$\chi_r^2 = \frac{12}{Nk(k + 1)} \Sigma(R_j)^2 - 3N(k + 1),$$

will be included in a future statistical article.

a mean rank within treatment within class. This score proved very useful, since it fulfilled certain requirements for parametric data.

Results

Comment vs. no comment. The over-all significance of the comment effects, as measured by the analysis of variance by ranks, is indicated in Table 2. The first row shows results obtained when students were considered as matched independently from one common population. The second row shows results when treatment groups within classes were regarded as intact groups. In either case the conclusions were the same. The Specified Comment group, which received automatic impersonal comments according to the letter grade received, achieved higher scores than the No Comment group. The Free Comment group, which received individualized comments from the teachers, achieved the highest scores of all. Not once in a hundred times would such differences have occurred by chance if scores were drawn from a common population. Therefore it may be held that the comments had a real and beneficial effect upon the students' mastery of subject matter in the various experimental classes.

It was also possible, as indicated earlier, to use the mean ranks within treatments

TABLE 2

The Friedman Test of the Over-all Treatment Effects

Units considered	N	F	S	*df*	χ_r^2	*p*
Individual Subjects	1363	1488	1427	2	10.9593	$< .01$
Class-group Subjects	129.5	170.0	144.5	2	11.3310	$< .01$

TABLE 3

Parametric Data Based Upon Mean Ranks Within Treatments Within Classes

Source	N	F	S	Total
Number of Groups	74	74	74	222
Sum of Mean Ranks	140.99	154.42	148.59	444.00
Sum of Squares of Mean Ranks	273.50	327.50	304.01	905.01
Mean of Mean Ranks	1.905	2.087	2.008	2.000
S.D. of Mean Ranks	.259	.265	.276	
S.E. of Mean Ranks	.030	.031	.032	

TABLE 4

Analysis of Variance of Main Treatment Effects

(Based on Mean Ranks)

Source	Sum of Squares	*df*	Mean Square	*F*	Probability
Between Treatments: N, F, S	1.23	2	.615	5.69	$< .01$
Between Class-groups	0.00	73	.000	. . .	
Interaction: T × Class	15.78	146	.108		
Total	17.01	221			

Note.—Modeled after Lindquist (3), p. 157 *et passim*, except for unusual conditions noted.

within classes as parametric scores. The resulting distributions, being normally distributed and fulfilling certain other assumptions underlying parametric tests, permitted other important comparisons to be made.[5] Table 3 shows the mean-ranks data necessary for such comparisons.

The various tests are summarized in Tables 4 and 5. The over-all F test in Table 5 duplicated, as one would expect, the result of the Friedman test, with differences between treatment groups still significant beyond the .01 level. Comparisons between different pairs of treatments are shown in Table 5. All differences were significant except that between Free Comment and Specified Comment. It was plain that comments, especially the individualized comments, had a marked effect upon student performance.

Comments and schools. One might question whether comment effects would vary from school to school, and even whether the school might not be the more appropriate unit of analysis. Since as it happened the study had 12 junior or senior high schools which had three or more experimental classes, these schools were arranged in a treatments-by-replications design. Results of the analysis are shown in Table 6. Schools apparently had little measurable influence over treatment effect.

Comments and school years. It was conceivable that students, with increasing age and grade-placement, might become increasingly independent of comments and other personal attentions from their teachers. To test such a belief, 66 class-groups, drawn from the experimental classes, were stratified into six school years (Grades 7–12) with 11 class-groups in each school year. Still using mean ranks as data, summations of such scores were as shown in Table 7. Rather surprisingly, no uniform trend was apparent. When the data were tested for interaction of school year and comment effect (see Table 8), school year did not exhibit a significant influence upon comment effect.

TABLE 5

DIFFERENCES BETWEEN MEANS OF THE TREATMENT GROUPS

Comparison	Difference	S. E. of Diff.	t	Probability
Between N and F	.182	.052	3.500	<.001
Between N and S	.103	.054	1.907	<.05
Between F and S	.079	.056	1.411	<.10(n.s.)

Note.—The t tests presented are those for matched pairs, consisting of the paired mean ranks of the treatment groups within the different classes. Probabilities quoted assume that one-tailed tests were appropriate.

Though Table 8 represents a comprehensive test of school-year effect, it was not supported by all available evidence. Certain other, more limited tests did show significant differences in school year, with possibly greater responsiveness in higher grades. The relevant data (**6**, chap. 5) are too cumbersome for the present report, and must be interpreted with caution. Apparently, however, comments do *not* lose effectiveness as students move through school. Rather they appear fairly important, especially when individualized, at all secondary levels.

One must remember that, between the present class-groupings, there were many differences other than school year alone. Other teachers, other subject-matter fields,

[5] It may be noted that the analysis of variance based upon such mean ranks will require no calculation of sums of squares between levels or between classes. This is true because the mean for any class will be $(k + 1)/2$, or in the present study just 2.00. . . . An alternative to such scores would be the conversion of all scores to T scores based upon each class-group's distribution; but the mean ranks, while very slightly less sensitive, are much simpler to compute and therefore less subject to error.

TABLE 6

The Influence of the School Upon the Treatment Effect

Source	Sum of Squares	*df*	Mean Square	*F*	Probability
Between Treatments: N, F, S	.172	2	.086	...[a]	...
Between Schools	.000	11	.000		
Between Classes Within Schools (pooled)	.000	24	.000		
Interaction: T × Schools	1.937	22	.088	...	...
Interaction: T × Cl. W. Sch. (pooled)	4.781	48	.099		
Total	6.890	107			

Note.—Modified for mean-rank data from Edwards (**1**, p. 295 *et passim*).

[a] Absence of an important main treatment effect is probably caused by necessary restriction of sample for school year (*N* is 36, as compared with Total *N* of 74), and by some chance biasing.

TABLE 7

Sums of Mean Ranks for Different School Years

School Year	N	F	S
12	21.08	22.92	22.00
11	19.06	23.91	23.03
10	20.08	23.32	22.60
9	22.34	22.06	21.60
8	21.21	22.39	22.40
7	22.04	22.98	20.98

Note.—Number of groups is 11 in each cell.

other class conditions could conceivably have been correlated beyond chance with school year. Such correlations would in some cases, possibly, tend to modify the *visible* school-year influence, so that illusions would be created. However possible, such a caution, at present, appears rather empty. In absence of contradictory evidence, it would seem reasonable to extrapolate the importance of comment to other years outside the secondary range. One might predict that comments would appear equally important if tested under comparable conditions in the early college years. Such a suggestion, in view of the large lecture halls and detached professors of higher education, would appear one of the more striking experimental results.

Comments and letter grades. In a questionnaire made out before the experiment, each teacher rated each student in his class with a number from 1 to 5, according to the student's *guessed responsiveness* to comments made by that teacher. Top rating, for example, was paired with the

TABLE 8

The Influence of School Year Upon Treatment Effect

Source	Sum of Squares	*df*	Mean Square	*F*	Probability
Between Treatments: N, F, S	1.06	2	.530	5.25	<.01
Between School Years	0.00	5	.000		
Between Cl. Within Sch. Yr. (pooled)	0.00	60	.000		
Interaction: T × School Year	1.13	10	.113	1.12	(n.s.)
Interaction: T × Class (pooled)	12.11	120	.101		
Total	14.30	197			

Note.—Modified for mean-rank data from Edwards (**1**, p. 295 *et passim*).

description: "Seems to respond quite unusually well to suggestions or comments made by the teacher of this class. Is quite apt to be influenced by praise, correction, etc." Bottom rating, on the other hand, implied: "Seems rather negativistic about suggestions made by the teacher. May be inclined more than most students to do the opposite from what the teacher urges." In daily practice, many teachers comment on some papers and not on others. Since teachers would presumably be more likely to comment on papers of those students they believed would respond positively, such ratings were an important experimental variable.

Whether teachers *were* able to predict responsiveness is a complicated question, not to be reported here. It was thought, however, that teachers might tend to believe their able students, their high achievers, were also their responsive students. A contingency table was therefore made, testing the relationship between *guessed* responsiveness and letter grade achieved on the first test. The results were as predicted. More "A" students were regarded as highly responsive to comments than were other letter grades; more "F" students were regarded as negativistic and unresponsive to comments than were other letter grades; and grades in between followed the same trend. The over-all C coefficient was .36, significant beyond the .001 level.[6] Plainly teachers believed that their *better* students were also their more *responsive* students.

If teachers were correct in their belief, one would expect in the present experiment greater comment effect for the better students than for the poorer ones. In fact, one might not be surprised if, among the "F" students, the No Comment group were even superior to the two comment groups.

TABLE 9

MEAN OF MEAN RANKS FOR DIFFERENT LETTER GRADES

Letter Grade	N	F	S
A	1.93	2.04	2.03
B	1.91	2.11	1.98
C	1.90	2.06	2.04
D	2.05	1.99	1.96
F	1.57	2.55	1.88

Note.—Each eligible class was assigned one mean rank for each cell of the table.

The various letter grades achieved mean scores as shown in Table 9, and the analysis of variance resulted as shown in Table 10. There was considerable interaction between letter grade and treatment effect, but it was caused almost entirely by the remarkable effect which comments appeared to have *on the "F" students*. None of the other differences, including the partial reversal of the "D" students, exceeded chance expectation.

These data do not, however, represent the total sample previously used, since the analysis could use only those student levels in which all three students received the same letter grade on Test One.[7] Therefore many class-groups were not represented at all in certain letter grades. For example, although over 10% of all letter grades were "F," only 28 class-groups had even one level consisting entirely of "F" grades, and most of these classes had *only* one such level. Such circumstances might cause a somewhat unstable or biased estimate of effect.

Within such limitations, the experiment

[6] In a 5 × 5 table, a perfect correlation expressed as C would be only about .9 (McNemar [4], p. 205).

[7] When levels consisted of both "A" and "B" students, for example, "A" students would tend to receive the higher scores on the second test, regardless of treatment; thus those Free Comment "A" students drawn from mixed levels would tend to appear (falsely) more responsive than the Free Comment "B" students drawn from mixed levels, etc. Therefore the total sample was considerably reduced for the letter-grade analysis.

TABLE 10

THE RELATION BETWEEN LETTER GRADE AND TREATMENT EFFECT

Source	Sum of Squares	*df*	Mean Square	*F*	Probability
Between Treatments: N, F, S	2.77	2	1.385	5.41	<.01
Between Letter Grades	0.00	4	0.000		
Bet. Blocks Within L. Gr. (pooled)	0.00	65	0.000		
Interaction: T × Letter Grades	4.88	8	.610	2.40	.05>p>.01
Residual (error term)	32.99	130	.254		
Total	40.64	209			

Note.—Modified for mean-rank data from Lindquist (3, p. 269). Because sampling was irregular (see text) all eligible classes were randomly assigned to 14 groupings. This was done arbitrarily to prevent vacant cells.

provided strong evidence against the teacher-myth about responsiveness and letter grades. The experimental teachers appeared plainly mistaken in their faith that their "A" students respond relatively brightly, and their "F" students only sluggishly or negatively to whatever encouragement they administer.

SUMMARY

Seventy-four randomly selected secondary teachers, using 2,139 unknowing students in their daily classes, performed the following experiment: They administered to all students whatever objective test would occur in the usual course of instruction. After scoring and grading the test papers in their customary way, and matching the students by performance, they randomly assigned the papers to one of three treatment groups. The No Comment group received no marks beyond those for grading. The Free Comment group received whatever comments the teachers felt were appropriate for the particular students and tests concerned. The Specified Comment group received certain uniform comments designated beforehand by the experimenter for all similar letter grades, and thought to be generally "encouraging." Teachers returned tests to students without any unusual attention. Then teachers reported scores achieved on the next objective test given in the class, and these scores became the criterion of comment effect, with the following results:

1. Free Comment students achieved higher scores than Specified Comment students, and Specified Comments did better than No Comments. All differences were significant except that between Free Comments and Specified Comments.

2. When samplings from 12 different schools were compared, no significant differences of comment effect appeared between schools.

3. When the class-groups from six different school years (grades 7–12) were compared, no *conclusive* differences of comment effect appeared between the years, but if anything senior high was more responsive than junior high. It would appear logical to generalize the experimental results, concerning the effectiveness of comment, at least to the early college years.

4. Although teachers believed that their better students were also much more responsive to teacher comments than their poorer students, there was no experimental support for this belief.

When the average secondary teacher takes the time and trouble to write comments (believed to be "encouraging") on student papers, these apparently have a measurable and potent effect upon student effort, or attention, or attitude, or whatever it is which causes learning to improve, and this effect does not appear dependent on school building, school year, or student ability. Such a finding would seem very important for the studies of classroom learning and teaching method.

REFERENCES

1. EDWARDS, A. *Experimental design in psychological research*. New York: Rinehart, 1950.
2. FRIEDMAN, M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J. Amer. statist. Ass.*, 1937, **32,** 675–701.
3. LINDQUIST, E. F. *Design and analysis of experiments in psychology and education*. Boston: Houghton Mifflin, 1953.
4. MCNEMAR, Q. *Psychological statistics*. (2nd ed.) New York: Wiley, 1955.
5. PAGE, E. B. Educational research: replicable *or* generalizable? *Phi Delta Kappan*, 1958, **39,** 302–304.
6. PAGE, E. B. The effects upon student achievement of written comments accompanying letter grades. Unpublished doctoral dissertation, Univer. of California, Los Angeles, 1958.

Received March 6, 1958.

Journal of Educational Psychology
Vol. 49, No. 4, 1958

COMPARISON OF ORGANISMIC AGE AND REGRESSION EQUATIONS IN PREDICTING ACHIEVEMENTS IN ELEMENTARY SCHOOL

HERBERT J. KLAUSMEIER, ALAN BEEMAN, AND IRVIN J. LEHMANN
University of Wisconsin

Olson (7) and Millard (6) report that both the average of and the relationships among seven ages in months—height, weight, grip, dental, carpal, mental, and reading—are useful in appraising the child's level of performance in other areas such as language and arithmetic. Klausmeier (5), however, found no statistically significant differences in height, weight, grip, dentition, and carpal age between high- and low-achievers in arithmetic and language. With this finding the present study was undertaken to ascertain how useful the first seven measures were (*a*) when combined in a best-combination regression equation and (*b*) when considered of equal weight as in Olson's system of organismic age, for predicting arithmetic and language achievement 12 months after the original measures were secured.

Procedure

The subjects of this investigation were 21 boys and 24 girls who were third-graders in 1955–56 and fourth-graders in 1956–57 and 29 boys and 24 girls who were fifth-graders in 1955–56 and sixth-graders in 1956–57. These children were enrolled in four regular classes of two large elementary schools of Madison and were all the children enrolled in the classes at the times the two sets of measures were secured.

The following measures were obtained: standing height, weight, strength of grip of the preferred hand, number of permanent teeth, bone development of the wrist and hand, mental age with the California Short-Form Test of Mental Maturity (S-Form), and achievement in reading, arithmetic, and language with the California Achievement Tests (Complete Battery, Form AA, Primary, Elementary, and Intermediate). In all instances results of a single measure were used except for strength of grip and carpal age. In securing strength of grip, a first measure was taken with the palm of the hand upward, grasping the dynamometer; the second with the palm downward; and the third with the palm upward. The average of the highest two of the three measures was used as strength of grip. This method was found to yield a higher test-retest correlation, 0.92, than using the single highest score with the palm upward, 0.83. All X-rays were read independently by the same two resident radiologists and the average of the two readings in months was used as carpal age.

The above measures were secured in October, 1955, and again in October, 1956, within the same week for each of the four classroom groups, and each measure at approximately the same hour of the day. Over this twelve-month interval, the measures were found to be relatively consistent as the correlation coefficients in Table 1 show. The primary level of the mental maturity and the achievement tests was used in 1955, the elementary level in 1956. The elementary level of the achievement battery was used in the fifth grade, the intermediate level in the sixth. No change was made in level of the mental maturity test.

Evidence of reliability and validity of the measures is now given since measurement is crucial in this study. In 1955, a random sample of 30 children was drawn from the total population of this study and remeasured within 24 hours after the first measuring of height, weight, and strength of grip. The test-retest correlations were .99 for height, .99 for weight,

TABLE 1
CORRELATIONS BETWEEN MEASURES OBTAINED ONE YEAR APART, IN OCTOBER 1955 AND OCTOBER 1956

	Third-Fourth Graders		Fifth-Sixth Graders	
	Boys	Girls	Boys	Girls
Height (inches)	.99	.82	.96	.95
Weight (pounds)	.87	.96	.97	.96
Grip (kilograms)	.64	.83	.76	.62
Permanent teeth	.47	.46	.77	.70
Carpal Age (month)	.85	.73	.83	.72
Mental Maturity Score	.50	.43	.76	.86
Reading Test Score	.55	.85	.86	.63
Arithmetic Score	.75	.86	.76	.78
Language Test Score	.87	.66	.62	.71

and .92 for grip. The two radiologists' independent readings of all the X-rays showed a correlation of .95 for third-graders and .86 for fifth-graders. Checks of successive dental records by the researchers showed that the dentist had identified permanent teeth without error. Thus, reliability of the five physical measures is considered high; and the test manuals report high reliability for the intelligence and achievement measures used. In addition, the correlations on the achievement measures reported above over the 12-month period indicate quite high reliability.

Concurrent validity of certain measures was also determined with the third-grade children. For a random sample of 30, California M.A. and Stanford-Binet M.A., obtained within six weeks of each other, correlated .82. Scores from the California Achievement Test in Reading correlated .90 with the Gates Advanced Primary Reading Test. Also, each of the four teachers ranked their children from highest to lowest in reading, arithmetic, and language achievement. The resulting rank-order correlations between test scores and teacher ratings are: reading—.91, .91, .90 and .88; arithmetic—.65, .82, .77 and .56; language—.69, .83, .85 and .77. In each set of four correlations, the first two are for third-grade and the last two for fifth-grade classes. Considering the difficulty of arranging a group of 30 children in rank order in each of the subject-matter areas, the researchers consider the correlations as indicating that the achievement tests measure the teachers' objectives sufficiently well for the purpose of this study.

FINDINGS

In Table 2 are presented the mean Pearson product-moment correlations among the measures as obtained in October, 1955. The mean correlations for the four groups of boys and girls are presented in Table 2 rather than each correlation on which the regression equations are based in

TABLE 2
MEAN CORRELATIONS AMONG RAW SCORES IN NINE MEASURES OF BOYS AND GIRLS, THIRD AND FIFTH GRADES

	Height	Weight	Grip	Dental	Carpal	Mental	Rdg.	Arith.	Lang.
Height		.65	.46	.00	.42	.05	.00	.01	.00
Weight			.28	.02	.35	−.01	−.04	−.01	−.02
Grip				.13	.39	.12	.15	.18	.02
Dental					.18	−.07	.05	.10	.04
Carpal						.05	.04	.09	.00
Mental							.62	.64	.59
Reading								.77	.75
Arithmetic									.74

TABLE 3

Contributions of Seven Measures to the Corrected Multiple *R*s and Beta Weights for the Regression Equation to Predict Arithmetic Achievement

	Height	Weight	Grip	Dental	Carpal	Mental	Reading
3rd boys							
Beta weight					.336		.819
R					.890 (2)		.830 (1)
3rd girls							
Beta weight				.219		.419	.534
R				.766 (3)		.734 (2)	.712 (1)
5th boys							
Beta weight	−.302	.371	−.258			.300	.423
R	.748 (5)	.722 (4)	.719 (3)			.695 (2)	.673 (1)
5th girls							
Beta weight			.210	.159		.423	.474
R			.890 (3)	.906 (4)		.886 (2)	.855 (1)

Note.—Blanks signify that the measure did not contribute .01 to the corrected Multiple *R* and did not then enter the regression equation.

order to present a more concise summary. For a correlation to be statistically significant from 0 at the .05 level (**2**), it must be between .367 and .404, depending upon the size of the *N* previously given. Table 2 shows that no mean correlation between the five physical measures and the three achievement measures is significant at the .05 level and dentition does not correlate significantly with any physical measure. Of the original 80 correlations between physical and achievement measures, only two, weight and language—fifth grade, were significant at the .05 level. However the other two measures comprising the basis of organismic age, mental and reading, correlate positively and significantly with arithmetic and language achievement.

Multiple correlations were computed, using the original correlations by grade and sex, and regression equations were derived to predict language and arithmetic achievement 12 months later. The multiple *R* and regression equation were calculated by the Wherry-Doolittle Test Selection Method (**3**). Any of the seven measures contributing .01 or more to the multiple *R*, uncorrected for shrinkage, were included in the multiple regression equation, provided their inclusion did not actually lower the multiple *R* when corrected for shrinkage. The Beta weights in the multiple regression equations for predicting scores were secured with the IBM 650 computer by a method of inverse correlation matrices. Table 3 shows the corrected multiple *R*s obtained between the seven organismic measures and arithmetic achievement, the order in which the various measures went into the multiple correlations, and the Beta weights for each regression equation. Reading correlated higher than any other measure with arithmetic for the four groups and thus went first into the multiple *R* and the regression equation. Differences between groups in the order in which the seven measures went into the regression equations and differences in Beta weights are not so important, however, as the finding that the best combination of all five physical measures increased the corrected multiple *R* by only .060 for the third-grade boys, by .032 for third-grade girls, by .053 for fifth-grade boys, and by .020 for fifth-grade girls. Table 4 shows similarly that the physical measures increased the corrected multiple

TABLE 4

Contributions of Seven Measures to the Corrected Multiple *R*s and Beta Weights for the Regression Equation to Predict Language Achievement

	Height	Weight	Grip	Dental	Carpal	Mental	Reading
3rd boys							
Beta weight							
R							.854(1)
3rd girls							
Beta weight		.317		.264	−.341	.317	.527
R		.766(3)		.788(5)	.759(2)	.767(4)	.746(1)
5th boys							
Beta weight		.406	−.460		.211	.297	.360
R		.826(4)	.764(3)		.845(5)	.662(1)	.718(2)
5th girls							
Beta weight	.209	−.422					.664
R	.763(3)	.762(2)					.729(1)

Note.—Blanks signify that the measure did not contribute .01 to the corrected Multiple *R* and did not then enter the regression equation.

R for language above that obtained with reading or a combination of reading and mental age by .00 for third-grade boys, .042 for third-grade girls, .127 for fifth-grade boys, and .034 for fifth-grade girls.

In an attempt to ascertain whether organismic age is a better predictor of achievement in arithmetic and language than is the predicted score derived by means of regression equations, Pearson product-moment correlations were calculated between regression-equation predicted scores and the raw scores in arithmetic and language obtained one year later, and also between organismic age in months and the scores obtained one year later. These results are presented in Table 5.

Table 5 indicates that the correlations are higher between the 1956 obtained and 1956 predicted scores derived from regression equations than between the 1956 obtained scores and the 1955 organismic age. Four of the eight correlations between organismic age and predicted achievement in arithmetic and language are significant at the .05 level; but all the correlations between regression-equation predicted scores and obtained scores are significant beyond the .01 level. It is anticipated, of course, that were the same regression equations applied to other samples, the obtained correlations between predicted and actual scores would be lower.

The present results are in accord with

TABLE 5

Correlations between Organismic Age Prediction, Regression-Equation Prediction and Obtained Arithmetic and Language Scores

Group	*N*	Arithmetic Score and Organismic Age Prediction	Language Score and Organismic Age Prediction	Arithmetic Score and Regression Prediction	Language Score and Regression Prediction
3rd boys	21	.370	.273	.683	.768
3rd girls	24	.039	.491	.731	.665
5th boys	29	.349	.472	.629	.657
5th girls	24	.586	.581	.743	.693

those of Gates (**4**) and Blommers, Knief and Stroud (**1**), who used designs different from the present study.

Summary

This study was conducted to compare the efficiency of organismic age, the average of seven ages, and regression equations, based on raw scores in the same seven measures, in predicting arithmetic and language achievements of third- and fifth-grade children 12 months after the original measures were secured. The regression-equation predictions correlated higher with actual achievements than did organismic-age predictions. The five physical measures in organismic age contributed little to mental and reading scores in predicting arithmetic and language scores.

References

1. Blommers, P., Knief, L. M., & Stroud, J. B. The organismic age concept. *J. educ. Psychol.*, 1955, **46,** 142–150.
2. Edwards, A. L. *Statistical methods for the behavioral sciences.* New York: Rinehart, 1954.
3. Garrett, H. E. *Statistics in psychology and education.* New York: Longmans, Green, 1953.
4. Gates, A. I. The nature and educational significance of physical status and of mental, physiological, social, and emotional maturity. *J. educ. Psychol.*, 1924, **15,** 329–358.
5. Klausmeier, H. J. Physical, behavioral, and other characteristics of high- and lower-achieving children in favored environments, *J. educ. Res.*, in press.
6. Millard, C. V. *Child growth and development.* Boston: Heath, 1951.
7. Olson, W. O. *Child development.* Boston: Heath, 1949.

Received January 25, 1958.

Journal of Educational Psychology
Vol. 49, No. 4, 1958

SEX DIFFERENCES IN THE RETENTION OF QUANTITATIVE INFORMATION

ROBERT SOMMER[1]

The Saskatchewan Hospital, Weyburn

Sex differences in arithmetic reasoning and spatial relations have consistently been found. Men are superior to women on these tests, while women excell on tasks requiring verbal ability and memory (**1, 5**). However, several authors (**1, 5**) state that the female superiority in recall does not hold if the material is more interesting to the males. This is a reasonable statement but there has been very little research designed to investigate the reasons for this. It is also interesting to note that there has been far more research relating to motivational factors in perception than there has been research relating to motivational factors in memory. Research on this latter point should be of concern to educators and others who hope to teach subject matter to students who vary widely in their interest in the content of courses.

This study had its origins in the observation that during routine testing of hospital patients with the Wechsler-Bellevue Information Scale, men did consistently better than women on items involving estimations of size or distance. Even intelligent women usually would be unaware of the population of the United States or the distance from New York to Paris. There seemed an inability to retain this information. It should be stressed that this was not a matter of computation or of analytic reasoning, rather it was a question of the retention of information that they had been exposed to a number of times.

This sex difference is not unknown to writers and educators. Weber (**6**) writes, "Mention mathematics to a women and she freezes into a condescending attitude of tolerance—she knows it exists, she uses it when she must, but it certainly has very little to do with her own delightfully imaginative and delicate world of interests." However our experience has been that this debility is more fundamental than simply a reflection of hostility to mathematical reasoning. Schilder (**3**) speaks of our remembering "only what we can and will use in the present situation." If this is so, then investigation into the retention of information should lead us into some rather basic attitudes regarding what type of information women believe is useful.

The purpose of the present studies is to determine whether these sex differences in recall of sizes and distances, observed clinically with hospital patients, would appear in other samples. Also of interest is whether there will be sex differences in the immediate recall of new quantitative material.

Experiment One

Procedure

Two populations of *S*s were sampled. The first consisted of patients in a mental hospital whose psychological test records were available in the files. The second was composed of students in elementary psychology classes at a small Midwestern college. The patient sample included only those whose last initials began with B, N, P, and T, who had been tested with the Information Scale of the Weschler-Bellevue, who were between 18 and 69 years old, and who had IQ's of above 70. This provided 156 cases, 96 males and 60 fe-

[1] This study was started by the writer and John Hinkle, Prabha Khanna, and Walter McDonald, and carried to conclusion by the present writer. We would like to thank J. B. Ray, G. McMurray, and H. Cooperstock for use of their classes for testing.

TABLE 1

PERCENTAGE OF CORRECT RESPONSES TO QUANTITATIVE ITEMS

Item	Male patients $N = 96$	Female patients $N = 60$	p**	Male students $N = 61$	Female students $N = 34$	p
Population (U.S.)	41	18	(.01)	64	47	(.08)
Distance	36	10	(.001)	54	59	(N.S.)
Pints	*	*		72	91	(.02)
Teaspoons	*	*		13	70	(.001)
Population (college town)	*	*		51	28	(.05)

* Item either not administered or scored for this group

** All p values are based on chi-square tests.

males. The students were all between 18 and 28 years old and constituted a sample of 61 males and 34 females.

The procedure for the patients was simply to tally and compare the number of correct "population of U.S." and "distance from New York to Paris" responses from males and females. (It can be noted that there was not a significant difference between the mean IQ of the males, 94.3, and that of the females, 95.6). The students were all tested in their classrooms by an examiner who requested them to answer the following questions:

1. How far is it from New York to Paris?

2. What is the population of the United States?

3. How many pints are there in a quart?

4. How many teaspoons are there in a tablespoon?

5. What is the population of (the town in which the college is located)?

Results

The responses to the information items are presented in Table 1. It is evident that there are very large differences in regards to estimating the population of the U.S. (with males excelling) and in recalling the number of teaspoons in a tablespoon (with females excelling) while many of the other differences are moderately large.

EXPERIMENT TWO

In a study of this sort where a classification of *S*s by sex is used, it is hazardous to speak of a genuine sex difference until a number of different groups have been sampled. Although the preceding table compared both college students and patients, it seemed in order to replicate the study with a fresh sample.

This time a group of 154 Canadian university students was used. The procedure followed was the same as in the previous study except that the questions were altered to suit the Canadian culture; that is, the *S*s were asked the population of Canada, the population of their university town, etc. The results are presented in Table 2. It is clear that the sex differences

TABLE 2

PERCENTAGE OF CORRECT RESPONSES TO QUANTITATIVE ITEMS

Item	Male students $N = 89$	Female students $N = 65$	p
Population (Canada)	85	55	.01
Distance	54	25	.01
Pints	83	86	N.S.
Teaspoons	20	48	.01
Population (university town)	89	52	.01

in the previous sample are supported by these results. The males do far better on the population and distance items, while the females excel in recalling the number of teaspoons in a tablespoon. There is no difference between the sexes in answering the pints in a quart item.

Experiment Three

With the procedures used in Experiments One and Two, we were not able to control the exposure of our *S*s to the information that was requested. Hence it was thought desirable to present new material to a group of *S*s and see if the males would surpass the females in remembering quantitative material.

Procedure

Two brief paragraphs were constructed, each containing both quantitative and nonquantitative material. Care was taken to see that the quantitative material should be "new" to the *S*s so that any differences could not be attributed to previous exposure. Hence the "facts" that were presented were fabricated and for the most part incorrect. The paragraphs were as follows:

> The Swedish ship, the Queen Fredrika, delivered its cargo of 12,000 pounds of wheat to Bombay. This city of 1,500,000 in a country of 264 million people is one of the richest trading ports in the Far East. The Captain of the ship was Olaf Hansen.

> Last week was the scene of a bloody revolution in Venezula. This country of 116,000 square miles is one of the richest oil-producing centers in the world. More than 1,200,000 barrels are shipped every month. The other important exports are tin, bananas, and cocoa.

The *S*s in the present study were 49 women and 27 men who were studying to be psychiatric nurses at a large mental hospital. All *S*s had at least an 11th grade education. The Director of Nurses Training reported that he did not feel that there was any difference between the men and the women in intelligence, except that the women "did better on the exams" than the men. However a control for IQ can be found in the number of items of nonquantitative material retained by the men and the women.

The *S*s were tested at their customary class sessions and were told that a paragraph would be read to them. When the examiner directed them to begin, they should write down all they remembered of it. The instruction to begin writing followed several seconds after the reading of each paragraph.

In scoring the recall data, the paragraphs were divided into "sense units" (similar in form to the units of the Wechsler Memory Scale). For example, /the Swedish ship/ the Queen Fredrika/ delivered its cargo/ of 12,000 pounds/ . . . etc., this yielded a total of 25 nonquanti-

TABLE 3

Average Number of Items Remembered (Nurses)

	Nonquantitative items				Quantitative items			
	Avg. No. Remembered	*SD*	*t*	*p*	Avg. No. Remembered	*SD*	*t*	*p* (one-tail)
Males $N = 27$	13.44	.72	.22	N.S.	1.30	.26	1.77	.05
Females $N = 49$	13.24	.58			.79	.12		

tative units and 5 quantitative units. A quantitative unit was scored as correct if the *number* was recalled correctly regardless of whether the unit (pounds, bushels, etc.) was accurate. All information as to the sex of the respondents was removed and the scoring was done by the writer. Twenty of the protocols were also scored independently by another researcher. The coefficient of agreement between scorers was .93 for the nonquantitative scores and 1.00 for the quantitative scores.

Results

The average number of quantitative items recalled by the men was 1.30 ± .26 while the average number recalled by the women was .79 ± .12. This difference is significant by t test at beyond the .05 level. On the nonquantitative items, no difference in recall was expected. As is shown in Table 3, this prediction was also confirmed.

As the level of significance of the difference in Table 3 is not high, it was decided to repeat the procedure using a fresh sample. The paragraphs were read to students in two elementary sociology classes at the University of Saskatchewan. The procedure was identical to that used with the nurses. The results are presented in Table 4 and show that the female students do slightly better than the males in recalling the nonquantitative information, but the males do significantly better than the females in recalling the quantitative information. When the two samples are pooled, a chi-square test shows that the sex difference in recalling the quantitative items is significant beyond the .01 level ($\chi^2 = 5.86$, $p < .01$).

Discussion

The results from Experiments One and Two confirm the prediction that female *S*s would be poorer than male *S*s on the two Wechsler Information items (population and distance) under consideration. It should be remembered that although these were designated as "quantitative items," they did *not* involve computation, judgement, or even analytic reasoning. To answer a question dealing with the population of the U.S. is not ordinarily a test in estimating size or number. No one has "seen" the population of the U.S. and few *S*s will attempt to find the answer by dividing the world's population by a set percentage. This item involves simply the retention of a word or number that one has seen and heard many times.

One can attempt to explain these results on the basis of the greater familiarity of the male *S*s with population and distance judgments and the females with pints and teaspoons. Yet this does not provide the whole answer for it is appar-

TABLE 4

Average Number of Items Remembered (Students)

	Nonquantitative items				Quantitative items			
	Avg. No. Remembered	*SD*	*t*	*p*	Avg. No. Remembered	*SD*	*t*	*p* (one-tail)
Males $N = 36$	15.06	.41	.81	N.S.	1.42	.69	1.73	<.05
Females $N = 74$	15.50	.36			1.00	.33		

ent that both men and women have been exposed to all of these items a number of times, surely enough for learning. We are all familiar with people who know the number of feet in a mile and the formula $E = mC^2$ but are unable to remember the number of pints in a quart. Such forgetting is highly selective as was shown in the Levine and Murphy experiment (**2**).

Experiment Three shows this difference appears with new material and cannot be attributed solely to a difference in previous contact with the information. This should have implications for the teaching of mathematics and other subjects to female students. Apparently the poorer performance of female students on tests of mathematics is more fundamental than a distaste for computation or algebra. Further research is necessary to determine the extent of this debility. There are some clues in our research that this deficiency is not directly related to an antipathy to all numbers. The females excelled in recalling the number of pints in a quart and teaspoons in a tablespoon. We also administered a brief test of digit span to the university students used in Experiment Three. Although the males had surpassed the females in recalling the quantitative information from the paragraphs, there was no difference in the recall of six and seven digits. This result parallels the negligible sex differences in recall of digits mentioned by Terman (**4**). Perhaps this indicates that many women are unable to retain *large* numbers (thousands or millions). An analysis was made of the type of errors in recalling the numbers from the paragraphs. It was found that 36% of these errors were due to the incorrect placement of the significant figures. That is, the *S* wrote "1200" or "120,000" instead of "12,000"; or "16,000," "1,600,000" or 1,016,00" instead of "116,000." This type of error constituted 40% of the incorrect responses by females and 27% of the incorrect responses by males. This difference is not significant and it should be realized that these are percentages of the *incorrect responses* given by all *S*s. That is, it does not include *S*s who did not write any figure or who wrote the correct figure. However it does show that there is need for research on the types of numbers that can be handled by men and women. If women are able to remember seven digits but can *not* remember five or six digit numbers it is important to learn the psychological characteristics of numbers qua numbers, instead of numbers as unrelated series of digits.

Summary

This study was undertaken to determine whether some differences between male and female patients seen in a hospital setting in the retension of quantitative information would be found in further tests. Three groups of *S*s were used: 156 hospital patients, 95 U.S. college students, and 154 Canadian college students. They were given several Wechsler-Bellevue Information items (population of U.S., pints in a quart, distance from New York to Paris) and a few other items. The results disclosed that the males did better on the population and distance items while the females performed better on the pints and teaspoons item. It was also shown that males were better able to retain new quantitative information when tested for immediate recall. No sex differences were found in remembering nonquantitative material. A brief digit span test also disclosed no sex differences. The implications of this for research into selective retention were discussed.

REFERENCES

1. Anastasi, A., & Foley, J. P. *Differential psychology*. N.Y.: MacMillan, 1949.
2. Levine, J. M., & Murphy, G. The learning and forgetting of controversial material. *J. abnorm. soc. Psychol.*, 1943, **38**, 507–517

3. SCHILDER, P. In D. Rapaport (Ed.), *Organization and pathology of thought*. N.Y.: Columbia University Press, 1951.

4. TERMAN, L. M. Psychological sex differences. In L. Carmichael (Ed.), *Manual of child psychology*. N.Y.: Wiley, 1946.

5. TYLER, L. E. *The psychology of human differences*. N.Y.: Appleton-Century-Crofts, 1947.

6. WEBER, NAOMI. Science in books. *Saturday Review of Literature*, Nov. 2, 1957, p. 43.

Received March 10, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 4, 1958

RELIABILITY OF THE CHILDREN'S MANIFEST ANXIETY SCALE AT THE RURAL THIRD GRADE LEVEL[1]

HAROLD D. HOLLOWAY

University of Tennessee

The children's form of the manifest anxiety scale (CMAS) developed by Castaneda, McCandless, and Palermo (**1**) offers a new and much needed group administered criterion of child behavior at the lower school-age levels, specifically for the fourth, fifth, and sixth grades. Criteria of these kinds are understandably more scarce at the first three school grades, a fact which prompted the present study.

Test-retest reliability coefficients reported by Castaneda, et al. (**1**)[2] ranged between .70 and .94 for samples of fourth, fifth, and sixth graders on whom the CMAS was standardized. In a series of articles using the CMAS as a predictor, the same investigators reported significant relationships between anxiety levels and performance on various learning tasks (**2**, **7**), preponderantly negative relationships between anxiety and popularity (**6**), and evidence to suggest that anxiety is meaningfully related to school achievement and intelligence for certain grade levels (**5**).

The purpose of the present research was to repeat essentially the reliability and standardization study of Castaneda, et al. using a *third* grade rural level in contrast to their samples of fourth, fifth, and sixth graders enrolled in a city school system.

METHOD

Subjects

A total of 121 children,[3] 64 boys and 57 girls, enrolled in four third grade classrooms of two rural schools located in an East Tennessee county served as *S*s. Three classrooms were in one school, and the fourth in another school. The schools, separated by about five miles or less, served adjacent communities of less than 2500 population. Parents of the *S*s were from a generally low- to middle-socioeconomic stratum as judged by occupational data. The *S*s were about equally distributed as to number and sex within classrooms.

CMAS Description

The CMAS consists of 53 items.[4] Forty-two items were designated by Castaneda, et al. as "anxiety" items and formed the A scale (abbreviation imposed by present author); 11 items were ". . . designed to provide an index of the subject's tendency to falsify his responses to the anxiety items . . ." (**1**, p. 318) and were labelled by the test authors as the L scale. By definition, the higher the A scale score, the higher the anxiety; and the higher the L scale score, the greater the tendency to falsify responses to the A scale.

[1] The study was sponsored jointly by the Agricultural Experiment Station and the College of Home Economics, Department of Child Development and Family Relationships of the University of Tennessee.

The author wishes to express his gratitude to the superintendent, principals and third grade teachers of Farragut and Blue Grass Elementary Schools of Knox County for their cooperation in making the study possible. The author is also indebted to the following persons for suggestions concerning the manuscript: Mary E. Keister, Boyd R. McCandless, Alfred Castaneda, Ruth Highberger, and William O. Jenkins.

[2] Not to be confused with Ref. (**2**)—hereafter, Castaneda, et al. refers to (**1**).

[3] One female *S* not included in analyses due to absence during second test administration.

[4] The items and scoring procedures may be found in (**1**, pp. 318–319).

Procedure

Two major steps were taken to maximize reading ease and comprehension: (*a*) Instructions were altered slightly so the teacher could give the items orally as the *S* followed on his own copy—in the Castaneda, et al. study, each *S* read and marked the items by himself. (*b*) Items were triple-spaced and typewritten in capitals. The instructions used in the present study are reproduced as follows:

> TO BOYS AND GIRLS
>
> Follow each question carefully as I read it aloud to you. When I finish reading each question to you, put a circle around the word YES if you think it is true about you. Put a circle around the word NO if you think it is not true about you. Now let us begin.

The testing program was carried out during the second half of the 1956–57 school year. The retest interval was approximately one week—seven days for three groups and six days for the fourth.

Results and Discussion

Scores obtained on the four groups were combined to form a single sample for the analyses. Table 1 includes the respective means (*M*s) and standard deviations (*SD*s) for the first and second A scales (A_1 and A_2), and similarly, for the first and second L scales (L_1 and L_2). Additionally, Table 1 contains the test-retest coefficients of reliability (Pearson *r*) for both scales.

As the various results are described, they will be compared with the Castaneda, et al. study. An attempt has been made to restrict comparisons mainly to trends, because the two studies concerned different grade levels, slightly different administrative instructions, and necessarily different error term components in tests of significance. Also, Castaneda, et al. presented only initial test data.

A and L Scale Ms *and* SDs

From Table 1 it is important to indicate that: (*a*) For both A and L, the *M*s for girls (see Col. 3) were higher than the corresponding *M*s for boys (see Col. 1) on both initial and final tests; and (*b*) for each sex, the *M*s of A_2 and L_2 (see Rows 2 and 6) were less than the corresponding *M*s of A_1 and L_1 (see Rows 1 and 5). Analyses of variance (Sex X Test Order) for the A and L scores separately resulted in one significant (coefficient of risk, $p = .01$) main effect—the pooled *M* of L_1 (4.70) was significantly higher than the pooled *M* of L_2 (4.02). Both interactions and other main effects were nonsignificant.

Using Sex X Grade analyses of variance with A_1 and L_1 scores as separate criteria,

TABLE 1

A and L Scale Test-Retest Means, SDs, and Reliabilities (*r*) for Pooled and Separate Sexes

Item	Boys		Girls		Pooled—Sex	
	M	*SD*	*M*	*SD*	*M*	*SD*
A_1	18.86	7.91	19.37	7.67	19.10	7.80
A_2	17.95	8.75	18.75	9.22	18.33	8.99
Pooled—A_1–A_2	18.41	8.35	19.06	8.48		
Reliability	$r = .82$		$r = .71$		$r = .83$	
L_1	4.34	2.39	5.11	1.90	4.70	2.20
L_2	3.78	2.00	4.44	2.29	4.09	2.17
Pooled—L_1–L_2	4.06	2.22	4.77	2.13		
Reliability	$r = .65$		$r = .78$		$r = .70$	

Note.—*N*s: Boys = 64; Girls = 57. All *r*s significantly different from zero—coefficient of risk, $p = .01$.

Castaneda, et al. found that girls scored *significantly* higher than boys on both scales. Thus, the significant sex differences of the Castaneda, et al. study were confirmed in direction by results of the present study.

In terms of magnitude, the Castaneda, et al. *M*s and *SD*s of A_1, and the *SD*s of L_1, were very close to the corresponding values of the present study. However, the *M*s of L_1 of the present study were approximately twice the size of those reported by Castaneda, et al., a finding that suggests an interesting hypothesis for further study.

Recall from above that on the initial scales there were higher anxiety levels and more falsification than on the final scales (compare pooled *M*s in right-hand section of Table 1). Clinically speaking, it would seem logical to expect such a result, since there was perhaps some anxiety associated with taking the test itself the first time which tended to dissipate during the week prior to taking the test a second time. The L scale may be viewed then as having "played the role" of a defense against anxiety, so that more falsification occurred in the initial test, when *S*s were presumably more anxious, than during the second test when less anxiety about test-taking was operating.

A and L Scale Frequency Distributions

Frequency polygons, smoothed by the method of running averages (**3**, pp. 52–54), were plotted for each sex separately and for both sexes combined using the A_1 and L_1 scale data. In general, the resultant six curves were unimodal, approximately bell-shaped, and fairly symmetrical. Most curves tended to possess a very slight positive skew but generally less skew than curve data presented by Castaneda, et al. For the pooled A_1 sample of the present study: median = 18.44; twentieth percentile (P_{20}) = 12.80; and P_{80} = 26.20. For the pooled L_1 sample the same statistics were respectively: 4.71; 2.73; and 6.73. Curves were not plotted for the A_2 and L_2 data, but inspection of the frequency distributions revealed them to be very similar to the A_1 and L_1 data. Considering the frequency data in general, the findings of the present study and those of Castaneda, et al. were in strong agreement.

A and L Scale Test-Retest Reliabilities

The *r*s between A_1 and A_2 ($r_{A_1 \cdot A_2}$) and between L_1 and L_2 ($r_{L_1 \cdot L_2}$), for pooled and separate sexes, are shown in Table 1. Three noteworthy features of the reliability results were: (*a*) All *r*s differed significantly from zero, represented substantial to marked relationships, and were generally comparable in size (although slightly lower in the case of A) to those of Castaneda, et al. (*b*) Coefficient $r_{A_1 \cdot A_2}$ was higher for boys (.82) than for girls (.71), a trend consistent with the fourth grade sample of Castaneda, et al. but opposite to their fifth and sixth grade samples. These combined facts suggested the hypothesis that within the age ranges included by both studies, boys tend to become less consistent in their responses to the A scale than do girls, but at the same time, the responses of both sexes maintain a relatively high level of consistency. (*c*) Due to the lack of significant sex differences for both A and L in the analyses of variance, although retention of the hypothesis of no sex differences does not prove it, it seemed reasonable to regard the pooled sex *r*s as the best single reliability estimates, viz., $r_{A_1 \cdot A_2}$ = .83 and $r_{L_1 \cdot L_2}$ = .70. The corresponding single estimates reported by Castaneda, et al. were .90 and .70 respectively, thus the two studies agreed very closely in this respect.

Correlation between the A and L Scales

The assumptions are made that: (*a*) the L scale indicates the tendency for *S* to falsify answers to the A scale; and (*b*) an attempt to falsify could result in either a high or low A scale score. Ideally then, from the standpoint of measurement, *r*s between the two scales would be zero or thereabouts. Correlations were computed between A_1 and L_1 and between A_2 and L_2 for each sex separately and for both sexes combined. Respectively, the $r_{A_1 \cdot L_1}$ and $r_{A_2 \cdot L_2}$ coefficients for the pooled sample were .14 and .05; for boys, .15 and .09; and for girls, .11 and .01. None of the *r*s was significantly different from zero (coefficient of risk, $p = .01$). These *r*s were generally comparable to those of Castaneda, et al.—theirs were also nonsignificant ranging between −.11 and .22.

General Conclusion

The evidence obtained strongly supported the findings of the test constructors, Castaneda, et al., who standardized their items on fourth, fifth, and sixth grade children. The principal conclusion drawn from the present study was that the A and L scales can be reliably employed as criteria using third grade rural children taken from populations similar to the one included herein. Whether or not the items are related to other operationally defined concepts (validity) is a matter for empirical determination.

Summary

The main purpose was to obtain test-retest coefficients of reliability of the Children's Form of the Manifest Anxiety Scale (CMAS) on a sample of 121 third grade rural *S*s. The CMAS consisted of 42 anxiety items (A scale) and 11 falsification items (L scale). The scales were administered to four classrooms by the respective teachers. The principal results, which were compared to those found in the reliability study of Castaneda, McCandless, and Palermo, are listed as follows:

1. Pooled estimates of the reliabilitites (r) of the A and L scales were .83 and .70 respectively. Correlations between the A and L scales approached zero.

2. Girls scored higher than boys on both scales but not significantly.

3. The general findings gave substantial support to those of Castaneda, et al. The evidence indicated that the A and L scales were sufficiently reliable to be used as criterion measures for samples from populations similar to those employed in the study.

REFERENCES

1. Castaneda, A., McCandless, B. R., & Palermo, D. S. The children's form of the manifest anxiety scale. *Child develpm.*, 1956, **27**, 317–326.
2. Castaneda, A., Palermo, D. S., & McCandless, B. R. Complex learning and performance as a function of anxiety in children and task difficulty. *Child Develpm.*, 1956, **27**, 327–332.
3. Guilford, J. P. *Fundamental statistics in psychology and education.* New York: McGraw-Hill, 1950.
4. Lindquist, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.
5. McCandless, B. R., & Castaneda, A. Anxiety in children, school achievement and intelligence. *Child Develpm.*, 1956, **27**, 378–382.
6. McCandless, B. R., Castaneda, A., & Palermo, D. S. Anxiety in children and social status. *Child Develpm.*, 1956, **27**, 385–391.
7. Palermo, D. S., Castaneda, A., & McCandless, B. R. The relationship of anxiety in children to performance in a complex learning task. *Child Develpm.*, 1956, **27**, 333–337.

Received March 10, 1958.

Journal of Educational Psychology
Vol. 49, No. 4, 1958

EFFECT OF PERIODICAL SELF-EVALUATION ON STUDENT ACHIEVEMENT

HENRY J. DUEL

Directorate of Civilian Personnel, USAF Headquarters

Interest in the application of techniques of self-evaluation to education and training is a relatively recent development. While there has been an increasing number of articles which indicate favorable experience with self-evaluation, research evidence concerning its value is lacking. Russell (**4**), in a 1953 survey of research on self-evaluation reports a lack of scientific study of the values of self-evaluation. Symonds (**5**) also indicates there are few reports on research results. Rogers (**3**), however, reports favorable experiential results with self-evaluation as a mode of appraisal. Thus self-evaluation has some empirical support but experimental evidence of its value is meagre.

The study reported here was undertaken to determine if self-evaluation is of value in improving student achievement. In other words, does self-evaluation give the student a basis for improved functioning as a student?

The study was conducted in two Air Force Schools at Scott Air Force Base, Illinois. *S*s were Air Force enlisted students in electronic communications courses. Due to similarity of the students' background, age, living conditions, aptitudes and to close control of curriculum, teaching methods, training materials and aids, conditions were considered especially favorable for a controlled study.

In conducting the study, self-evaluation instruments developed by the experimenter in a previous study were used (**2**). In that study he concluded that in schools of this nature, students could reliably and validly evaluate gain in skills and knowledges achieved in a technical course of instruction.

The instrument (which for convenience was called the SET) was one which required the student to make an estimate of the level of skill or knowledge he possessed upon entrance into the course as well as the skill or knowledge he attained upon completion of the course. A sample item from the form is shown below:

How proficient are you in using a multimeter to measure output voltages and currents of a vacuum tube?

1	2 3	4 5	6 7	8 9
Not prepared to do the job without thorough training	Am familiar with the job but need considerably more training and practice	Can do the job with further OJT and close supervision and assistance	Can do the job if given adequate supervision	Can do the job with no supervision

The student responded to each item by marking the scale with a check mark (✓) to indicate the level of skill he thought he possessed at the beginning of the course. An X represented his estimated attainment at the end of the course.

Using the SET as a device for self-evaluation, an experiment was organized in each of two schools. In School A, approximately 100 cases were used as control. The test group, also 100 cases, used the above-described device to evaluate themselves at the end of each "test point" of instruction. One to three weeks elapsed between each test point. Thus each student in the test group evaluated his progress in the course every one, two or three weeks of the course. Each SET encompassed skills learned in the previous two or three weeks.

School A consisted of 20 weeks of instruction for six hours per day, five days a week. In this case the test was closely controlled by the experimenter so that all instructors administering the SET gave the same instructions and administered it in the same manner.

In School B approximately 75 cases were entered as a test group and a like number was used as control. In this school, the SET was administered by school personnel and was handled as a normal part of class activity. This was done for the purpose of determining the effect of using self-evaluation in a "normal" class situation and to determine if special conditions of the experiment may have had effect on achievement results.

Student achievement was evaluated by regular course tests, used as a basis for determining grades. These criterion measures were carefully constructed and validated. Split-half reliabilities ranged from .74 to .90. The criterion measures had curriculum validity through construction, since "test blueprints" were used to derive items directly from curricula materials. In addition each item was evaluated by at least three experts as to its relevance to the job for which the man was being trained. Other item data including discrimination and difficulty indexes were used in construction of the tests.

Results from School A were obtained

TABLE 1

Test Score Means and t Ratios for Test-Control Groups in School A

(N = 75)

	Raw test score means by check points Check point no.								
	1	2	3	4	5	6	7	8	9
Test group mean	31.1	34.6	19.4	17.8	39.8	23.1	22.1	30.4	36.5
Control group mean	29.1	31.0	17.7	16.2	37.5	21.0	21.7	29.2	36.0
t	1.61	2.64*	3.26**	2.78**	3.25**	2.64*	.95	1.31	.78

* Significant at .02 level.

** Significant at .01 level.

TABLE 2

Test Score Means and t Ratios for Test-Control Groups in School B

(N = 33)

	Raw test score means by check points Check point no.								
	1	2	3	4	5	6	7	8	9
Test group mean	23.5	21.3	21.6	18.8	22.9	22.6	23.3	23.2	24.3
Control group mean	21.8	21.0	20.2	17.4	22.6	17.9	18.9	18.9	22.3
	1.93	.28	1.43	1.88	.36	3.07*	3.45*	3.41*	2.70*

* Significant at .01 level.

from a total of 75 paired cases which remained from the original 100 paired cases. Others were lost due to elimination and other administrative problems. Students in test and control groups were paired by aptitude index based on a battery of tests given at induction centers.

Table 1 shows results of all measures given to both groups in School A.

All nine measures indicate a positive difference in favor of the test group. Five of the nine differences show a t which is significant at the one per cent or two per cent level. Thus evidence strongly favors the test group. If all measures are combined by means of the use of multiple critical ratio, as suggested by Chapin (**1**), an MCR of 6.40 is obtained indicating a highly significant difference in favor of the self-evaluating group.

Results obtained from 33 matched pairs in School B are shown in Table 2.

Results in School B where "normal" use of self-evaluation was attempted, also indicate a positive difference in favor of the test group on all measures. Measures number eight and nine were not used because of improper administration. Four of the nine differences show a t which is significant at the one per cent level. An MCR of 6.17 also indicates highly significant difference in favor of the self-evaluating group.

Summary and Conclusions

This study was accomplished to determine the effect of periodic self-evaluation on student achievement. Students in two military technical schools periodically evaluated their skill and knowledge during a course of instruction, under careful control of the experimenter in one school and as part of the "normal" class activities in the second. Achievement of test groups on regular school tests was compared with that of control groups which were matched by aptitude. The results favored the self-evaluation group, with multiple critical ratios being statistically significant in both schools. The results lead to the conclusion that in this particular situation students, given formal and periodic opportunities to evaluate themselves, can achieve to a greater degree than students not having such opportunity.

The study also raises several questions: Does a device such as the one used furnish additional motivation, sharpen perceptions of the objectives to be achieved, or result in better organization of previous learning on which future learning is based? These and similar questions should furnish a basis for future study in the area of self-evaluation.

REFERENCES

1. Chapin, F. S. *Experimental designs in sociological research.* New York: Harper, 1947.
2. Duel, H. J. A study of validity and reliability of student evaluation of training. Unpublished doctoral dissertation, Washington Univer., 1956.
3. Rogers, C. R. *Client-centered therapy.* Boston: Houghton Mifflin, 1954.
4. Russell, D. H. What does research say about self-evaluation. *J. educ. Res.*, 1953, **46**, 563–571.
5. Symonds, R. M. Pupil evaluation and self-evaluation. *Tchrs. Coll. Rec.*, 1952, **54**, 138–149.

Received March 15, 1958.

Journal of Educational Psychology
Vol. 49, No. 4, 1958

AN EXPERIMENTAL EVALUATION OF THE OPEN BOOK EXAMINATION

RICHARD A. KALISH[1]

University of Hawaii

Part of the contemporary pedagogical trial-and-error efforts to improve techniques of classroom testing has focused upon the use of the open book examination. In such an examination the student is allowed to make use of any materials at his disposal, including textbooks, lecture notes, and dictionaries, but does not obtain answers either directly or indirectly from other students.

Tussing (**2**) summarizes the various arguments for using the open book test as:

> 1. The test can be constructed and used in all the various forms that the traditional test can be used; 2. Much of the fear and emotional blocks encountered by the student is removed; 3. Emphasis is placed upon the practical problems and reasoning, and less emphasis is placed upon pure memory of facts and items; 4. Cheating with cribs and other devices is eliminated; 5. This approach is more adaptable to evaluating student attitudes and posing the question of what action should be taken on social issues.

Some of the arguments opposing the use of the open book should also be recognized; namely, (*a*) It is likely to reduce study by allowing some students to feel that the use of the book will enable them to "slide through" with a minimum of study; (*b*) There is some reason to believe that a certain amount of rote memory may bring about the overlearning so often necessary to a full understanding of a subject; (*c*) Note-passing and looking at the test paper of a nearby student is made easier in the confusion of looking through papers and books; (*d*) A more superficial knowledge of the material is encouraged.

Problem

The present study is an attempt to determine the equivalence of two approaches to the administration of examinations; namely, the conventional closed book versus the open book. The general hypotheses are: (*a*) The open book examination will lead to fewer student errors; (*b*) The open book examination will measure different abilities than those assessed by the closed book tests; and (*c*) There is no correlation between student ratings of the help received from open book examinations and their test scores.

The first hypothesis is based on the apparent truism that the opportunity to look up material at its source should provide greater accuracy of response than depending upon memory. The null hypothesis may be stated as follows: An experimental group, receiving an open book examination, will not differ significantly in terms of total errors from a control group which receives the same examination under the traditional method.

The second hypothesis is based on the assumption that certain individuals will do better work on a closed book test while others will do relatively better on an open book examination, the differences being functions of differential responses to the pressure of the examination situation, an altering of motivation in studying, the ability to make organized use of texts and notes, etc. Specifically, then, the null hypothesis will state: The correlation between two closed book examinations will not differ significantly from the correlation between an open book and a closed book examination, assuming the sets of examinations and testing conditions are equivalent in every way.

[1] The author wishes to express his thanks to W. E. Vinacke and John Digman for their help with the manuscript.

The third hypothesis was based primarily upon the "educated guess" of the investigator which, in turn, was based on casual observations of the grades of students taking open book tests. The null hypothesis states: There will be no difference in number of errors on open book examinations between groups of students who rate open book examinations as being helpful and those who rate them as being nonhelpful. In this case, it is predicted that the null hypothesis will be accepted.

Method

Subjects. The subjects were 158 students at the University of Hawaii, 85% women and 75% sophomores. Seventy-four of the students were enrolled in one section of child psychology, while the remaining 84 were enrolled in another section of the same course with the same instructor.

Examinations. Two mid-term examinations, approximately six weeks apart, were given to both sections. Each examination consisted of 50 questions, all multiple-choice items with five alternative responses, only one of which was acceptable as correct. About one half the items were based on the class lectures; the other half were drawn from the text. Included were items which were distinctly factual in nature, items which attempted to get at understanding of relationships, and items which measured an understanding of terminology. It is believed that the tests were fairly typical of college tests of the multiple-choice variety.

The students were allowed 50 minutes for the examination, from the time the tests were completely distributed until the time they were collected. The score was the total number of errors, a high score thus indicating a low grade.

Procedure. The two sections, meeting in the same room at successive hours, were given the same six-week examinations on the same day at successive hours. Both sections had covered the same material during the class periods, had the same assignments, and received the same examinations. Opportunity for communication between sections was eliminated by keeping the students from the first section in the classroom until the end of the examination period, then releasing them by a back door while ushering the second class in through the front. There was little if any possibility for passing on questions and answers.

The first section to meet (Class A) was given the usual closed book examination for both six-week examinations. The second section (Class B) was given the same examinations, but only the first examination was in normal closed book form. At the first class meeting following the examination it was announced for the first time that the next examination would be "open book." Class B's second examination was taken with the use of textbooks, notes, and dictionaries. Otherwise, testing conditions were exactly the same as for Class A.

Most students in Class A had from 15 to 20 minutes remaining after completing the second examination, so it was assumed that most students in Class B should have had approximately that much time to look for answers in the material available to them.

At the close of the examination period, Class B was asked to indicate how much help the open book procedure provided by writing "None," "Little," "Some," or "Much" on their answer sheet.

Replication. A replication, comparable in every way except one, was conducted with 161 students, divided into Class A′ ($N = 79$) and Class B′ ($N = 82$). The one way in which the replication differed from the original study was that the two classes were not held in the same classroom and that communication between the two groups was not so well controlled. It is still highly unlikely that communication occurred to the extent of influencing the study.

Results

To test the first null hypothesis, it was first necessary to determine whether the two sections (Class A and Class B in the original experiment, Class A′ and Class B′ in the replication) were comparable in ability. This was accomplished by comparing their scores on the first examination, taken by both groups under the same conditions. Since the first and the second examinations were not necessarily of equal difficulty, the effects of the open book examination had to be measured in terms of the differences between the sections on the two examinations. These data are contained in Table 1.

As can be observed in Table 1, the scores were relatively the same for Class A and Class B (and for Class A′ and Class B′) on the first and second examinations. Although in each case the Experimental Group obtained approximately a one-half point relative increase under the experimental conditions, the difference is far from statistically significant.

Therefore, Null Hypothesis 1 must be accepted. It would appear from the results that, under the given conditions, the opportunity to use text and lecture materials resulted in no difference in total errors.

To test the second null hypothesis, Pearson product-moment correlations were computed for each class. The significance of the difference between correlations for Class A and Class B (and for Class A′ and Class B′) was then computed. The obtained correlations and their significance of difference levels are shown in Table 2. In both the original experiment and the replication, the correlation for the control condition was substantially higher than for the experimental condition (r's = .691 and .579 as opposed to .495 and .460), which was as hypothesized.

Although neither difference was significant with the two-tailed t test (t = 1.90 and 1.00), both were in the expected direction. The probabilities of the two experiments were combined according to the chi square method for independent samples suggested by Gordon, Loveland, and Cureton (**1**). The obtained chi square was 11.841, which is significant beyond the .02 level of confidence with four degrees of freedom.

TABLE 1

Mean Number of Errors on Examinations

	Mean Number Errors Both Closed Book (Exam I)	Mean Number Errors Experimental Condition (Exam II)	Change
Experiment			
Control (Class A)	10.24	15.68	+5.44
Experimental (Class B)	8.23	13.10	+4.87
Difference	2.01	2.58	0.57
Replication			
Control (Class A′)	13.51	13.41	−0.10
Experimental (Class B′)	15.56	14.88	−0.68
Difference	2.05	1.47	0.58

TABLE 2

Correlations Between Scores on First Examination and Second Examination and Tests of Significance Between Correlations

	Correlation	*t* Between groups	Level of confidence
Experiment			
Control (Class A)	.691		
Experimental (Class B)	.495	1.90	.056
Replication			
Control (Class A′)	.579		
Experimental (Class B′)	.460	1.00	.316
Combined $\chi_2 = 11.841$ $P < .02$			

TABLE 3

Change in Number of Errors From First Examination to Second Examination for the Experimental Groups as a Function of Attitudes Regarding Open Book Methods

	Extent of Help Received From Open Book Examination							
	N	Much help change	*N*	Some help change	*N*	Little help change	*N*	No help change
Class B	6	+5.16	33	+4.74	26	+4.96	4	+5.50
Class B′	9	−1.78	42	+0.12	15	−0.53	1	+2.00

Therefore, the second null hypothesis was rejected. It appears that a significantly lower correlation is obtained when an open book examination is given following a closed book examination than when both examinations are the closed book type.

For the final null hypothesis, the students were asked to indicate whether they felt the open book examination had been of "Much," "Some," "Little," or "No" help. As may readily be observed in Table 3, there was virtually no difference among the four groups of students indicating the four attitudes towards open book tests.

Not only are the differences among the groups slight, but for the first study, those who felt the open book was "Little" help did relatively better than those who felt it was "Much" help; in the replication a similar reversal occurred between "Some" help and "Little" help.

Therefore, the third null hypothesis was accepted, which was in accordance with the corresponding general hypothesis. It appears that the feelings of students regarding help given by an open book examination are not reflected in measured grade changes.

Discussion

This study has investigated the equivalence of the open book examination and the closed book examination. The results have indicated that, although under the conditions of this experiment the group average scores are not affected by the examination approach, the two types of examinations measure significantly different abilities.

While recognizing the obvious dangers

of over-generalizing, it is felt by the author that the experimental situation is sufficiently "typical" of college e:aminations of the multiple-choice variety to have widespread applicability.

It can be assumed, therefore, that some students do relatively better on open book examinations, while others do better on closed book examinations. Use of the open book technique, thus, would appear more rewarding to certain students than to others, and the belief that there is no difference between the two types of examination is of dubious validity.

Therefore, if an instructor feels, with Tussing, that the open book examination is a more valid measure due to the decrease in reliance on memory and detraction from cheating, the open book approach would be most appropriate. On the other hand, if he feels that the closed book type provides more study motivation and encourages a less superficial approach to a course, he will undoubtedly adhere to the traditional examination.

This study has shown that the two types of examinations measure significantly different abilities. It will now be necessary to investigate what factors differentiate students who are successful on each of the types of tests, so that instructor decisions might be based on more complete information.

Summary

An investigation was made of the equivalence between open book examinations and closed book examinations.

Two traditional closed book examinations were administered to a class of University of Hawaii students; the same examinations were administered to another section of students taking the same course with the same instructor, differing only in that the second examination was open book. A replication is also reported.

Three hypotheses were tested: 1. The open book examination will lead to fewer student errors; 2. The open book examination measures different abilities than the closed book examination; 3. Student ratings of the help received from open book examinations will not be related to examination scores. The first hypothesis was not substantiated, but the second and third hypotheses were verified.

REFERENCES

1. Gordon, M. H., Loveland, E. H., & Cureton, E. E. An extended table of chi-square for two degrees of freedom, for use in combining probabilities from independent samples. *Psychometrika*, **17**, 1952, 311–316.

2. Tussing, L. A consideration of the open book examination. *Educ. Psychol. Measmnt.*, **11**, 1951, 597–602.

Received March 28, 1958.

Journal of Educational Psychology
Vol. 49, No. 4, 1958

THE STANDARD ERROR OF MEASUREMENT OF THE DIFFERENCE BETWEEN A SUM SCORE AND ONE OF ITS PARTS

FREDERICK B. DAVIS

Hunter College

For proper interpretation of an individual's test scores it is sometimes necessary to ascertain the significance of the difference between a total score, consisting of the sum of several part scores, and one of those part scores. For example, the Level-of-Comprehension score from the Davis Reading Test (2) is based on the first 40 of the 80 items that determine the Speed-of-Comprehension score. A difference between individual speed and level scores should be evaluated in terms of the standard error of measurement of the difference between a sum (the Speed-of-Comprehension score) and one of its parts (the Level-of-Comprehension score).

For the convenience of clinical and school psychologists, the equations for computing the standard error of a difference between overlapping total and part scores obtained by an individual drawn at random from a specified group will be presented first and their use illustrated with data from the Davis Reading Test. The derivation of these new equations will then be provided.

Practical Procedures

Let T represent an individual's raw score made up of m parts. Then $T = A + B + \cdots + M$. Let I represent any part of sum T, and P any part of sum T except I. Differences between sum T and any one of its parts are inconvenient to interpret unless all of the scores are made comparable. For purposes of this discussion, comparable scores are defined as transformed raw-score values for which the corresponding true-score points are exceeded by the same percentage of examinees in a defined sample. The desired raw-score values may be determined by the method given by Flanagan (3, pp. 752–760). They are transformed simply to make numerically identical comparable scores for which corresponding true-score points are exceeded by the same percentage of examinees in the defined sample. For example, for Form A of the Davis Reading Test a speed score of 31 and a level score of 21 are raw-score values for which the corresponding true-score points are exceeded by about 61 per cent of the exaaminees in the equating sample (which comprised 4,692 students in Grades 11 and 12 and the freshman year of college). These two raw-score values have been transformed into comparable scores of 75. It should be made clear that comparable scores, as defined above, are not necessarily measures of the same abilities or equally reliable.

Fortunately, total and part scores are often expressed in serviceable approximations to comparable scores. For example, Verbal, Performance, and Full-Scale IQ's from the Wechsler intelligence scales are expressed in units such that their means are approximately 100, their standard deviations about 15, and the shapes of their distributions nearly normal. Similarly, total and part scores from the Cooperative Achievement Tests are expressed in Scaled Scores that have means and standard deviations (in a defined hypothetical group) of 50 and 10, respectively, and distributions that are closely normal.

Suppose that the raw scores (T, A, B, $\cdots$, M, as well as I and P) of the individual mentioned above are expressed in comparable form and denoted Z_T, Z_A, Z_B, $\cdots$, Z_M; and that Z_I denotes a comparable score on any part of sum T and Z_P any comparable part score except Z_I. Then it should be noted that $Z_T \neq Z_A + Z_B + \cdots + Z_M$. The standard error of measurement of the difference between Z_T and Z_I may be written as:

$$s_{meas\,(Z_T - Z_I)} = \frac{s_{(Z_T - Z_I)}}{s_{(T-I)}} \sqrt{\sum^{m-1} s_P^2(1 - r_{PP'})} \qquad [1]$$

or,

$$s_{meas(Z_T - Z_I)} = \frac{s_{(Z_T - Z_I)}}{s_{(T-I)}} \sqrt{\sum^{m-1} \left(\frac{s_P^2 \, s_{meas\,Z_P}^2}{s_{Z_P}^2}\right)} \qquad [2]$$

or,

$$s_{meas(Z_T - Z_I)} = s_{(Z_T - Z_I)} \sqrt{1 - r_{(T-I)(T'-I')}} \qquad [3]$$

where:

$$s_{(Z_T - Z_I)} = \sqrt{s_{Z_T}^2 + s_{Z_I}^2 - 2s_{Z_T} s_{Z_I} r_{TI}} \qquad [4]$$

$$s_{(T-I)} = \sqrt{s_T^2 + s_I^2 - 2s_T s_I r_{TI}} \qquad [5]$$

$$r_{(T-I)(T'-I')} = \frac{s_T^2 r_{TT'} + s_I^2 r_{II'} - 2s_T s_I r_{TI'}}{s_T^2 + s_I^2 - 2s_T s_I r_{TI}} \qquad [6]$$

and

$$r_{TI'} = r_{TI} - \frac{s_I(1 - r_{II'})}{s_T} \qquad [7]$$

In the preceding equations, s_T^2, s_I^2, and s_P^2 and $r_{TT'}$, $r_{II'}$, and $r_{PP'}$ are the variances and reliability coefficients, respectively, of these variables in the original raw-score units of measurement. The correlation of sum scores and any given set of part scores, expressed in these original units of measurement, is denoted as r_{TI}. Variances of the transformed comparable scores are denoted as s_Z^2, $s_{Z_I}^2$, and $s_{Z_P}^2$.

Whether the difference $Z_T - Z_I$ for any pupil chosen at random from the group tested may be regarded as a chance deviation from a true difference of zero at any desired level of confidence may be determined with serviceable accuracy by means of the critical ratio:

$$CR = \frac{(Z_T - Z_I) - 0}{s_{meas(Z_T - Z_I)}} \qquad [8]$$

Choice among Equations [1], [2], and [3] for computing the standard error of measurement of a difference depends on which one can be employed most conveniently with the data available. To determine the standard error of measurement of the difference between the speed and level scores from the Davis Reading Test for a college freshman drawn at random from a group tested, Equation [1] is most convenient. The test results give the standard errors of measurement of these scores, in terms of the original raw-score units of measurement, as 5.5 for speed and 3.7 for level. Equation [12], therefore, yields $(5.5)^2 - (3.7)^2$, or 16.56, as the numerical value under the radical sign in Equation [1]. Numerical values of the terms in Equations [4] and [5], also given in the manual, lead to a value of .35 for the ratio of $s_{(Z_T - Z_I)}$ to $s_{(T-I)}$. The standard error of measurement of the difference turns out to be 1.4. When this value is used in Equation [8], a difference of 2 points is found to be significant at about the 15 per cent level and one of 3 points at about the 3 per cent level. For a college student drawn at random from the group tested, a counselor or teacher would be justified in concluding that a difference between his speed and level scores of 3 points or more should be attributed to causes other than chance. It may occasion some surprise to find that the standard error of measurement of the difference between these two comparable scores is so small. This is largely accounted for by the fact that errors of measurement in them are positively correlated.

Derivation of Equations

Let T represent a raw total score made up of m parts: A, B, $\cdots$, M. Then $T = A + B + \cdots + M$. Let I represent any part of T and let P represent any part of T except I. Assume that an indefinitely large number of parallel forms of the tests from which these raw scores are derived are given to a pupil drawn at random from a grade group for which the tests are appropriate and postulate that this pupil's true scores in the abilities tested remain constant throughout the testing. An essentially normal distribution of differences between T and I would then be obtained. The mean of the distribution would approach $T_t - I_t$ (the difference between the pupil's true scores), and its variance could be written as:

$$\begin{aligned} s_{(T-I)}{}^2 = s_{(T_t+T_e)-(I_t+I_e)}{}^2 &= s_{T_t}{}^2 + s_{I_t}{}^2 \\ &+ s_{T_e}{}^2 + s_{I_e}{}^2 - 2s_{T_t}s_{I_t}r_{T_tI_t} \\ &+ 2s_{T_t}s_{T_e}r_{T_tT_e} + 2s_{I_t}s_{I_e}r_{I_tI_e} \qquad [9] \\ &- 2s_{T_t}s_{I_e}r_{T_tI_e} - 2s_{T_e}s_{I_t}r_{T_eI_t} \\ &- 2s_{T_e}s_{I_e}r_{T_eI_e} \end{aligned}$$

where the subscript t denotes a true score and the subscript e an error of measurement.

Since we postulated that the pupil's true scores remain constant, $s_{T_t}{}^2$ is equal to $s_{I_t}{}^2$ is equal to zero. Consequently, Equation [9] may be simplified to:

$$s_{meas_{(T-I)}}{}^2 = s_{T_e}{}^2 + s_{I_e}{}^2 - 2s_{T_e}s_{I_e}r_{T_eI_e} \qquad [10]$$

It can easily be shown that the coefficient

$$r_{T_eI_e} = r_{(T-T_t)(I-I_t)} = \frac{s_{I_e}}{s_{T_e}} \qquad [11]$$

By definition, $s_{T_e}{}^2$ is equal to $s_{meas_T}{}^2$ and $s_{I_e}{}^2$ is equal to $s_{meas_I}{}^2$. Therefore, Equation [10] may be simplified to:

$$s_{meas_{(T-I)}}{}^2 = s_{meas_T}{}^2 - s_{meas_I}{}^2 \qquad [12]$$

If we make the usual assumption that the correlation of errors of measurement of separate tests will, under proper conditions of test administration, be zero, we may write:

$$\begin{aligned} s_{meas_T}{}^2 = s_{T_e}{}^2 &= s_{(A_e+B_e+\cdots+M_e)}{}^2 \\ &= s_{A_e}{}^2 + s_{B_e}{}^2 + \cdots + s_{M_e}{}^2 + 0 \qquad [13] \\ &= \sum^{m-1} s_{meas_P}{}^2 + s_{meas_I}{}^2 \end{aligned}$$

If a substitution is made for $s_{meas_T}{}^2$ in Equation [12], we obtain:

$$\begin{aligned} s_{meas_{(T-I)}}{}^2 = \sum^{m-1} & s_{meas_P}{}^2 + s_{meas_I}{}^2 \\ & - s_{meas_I}{}^2 = \sum^{m-1} s_{meas_P}{}^2 \end{aligned} \qquad [14]$$

If sum T and each of its parts are transformed into comparable scores, as defined previously, we obtain Equations [1] and [2] by multiplying each side of Equation [14] by

$$\frac{s_{(Z_T-Z_I)}{}^2}{s_{(T-I)}{}^2}$$

and substituting

$$\frac{s_P{}^2}{s_{Z_P}{}^2}\, s_{meas_{Z_P}}{}^2$$

for $s_{meas_P}{}^2$.

Equation [3] is a specific application of the well-known relationship:

$$s_{meas_X} = s_X\sqrt{1 - r_{XX'}}$$

Equations [4], [5], and [6] are well known and the derivation of Equation [7] has been published by the writer (1).

REFERENCES

1. Davis, F. B. Note on part-whole correlation. *J. educ. Psychol.*, 1958, 49, 77–79.

2. Davis, F. B., & Davis, C. C. *Davis Reading Test.* Series 1, for High School and College Students. Forms A, B, C, and D. New York: Psychological Corp., 1957.

3. Flanagan, J. C. Units, scales, and norms. In E. F. Lindquist (Ed.), *Educational measurement.* Washington: American Council on Education, 1951.

Journal of Educational Psychology
Vol. 49, No. 4, 1958

A NOTE ON SEX EQUALITY IN THE INCIDENCE OF LEFT-HANDEDNESS[1]

WAYNE DENNIS
Brooklyn College

It is commonly accepted that comparisons of behavior in different cultures may provide data which are decisive for psychological theories. Yet the number of cross-cultural studies which have been used to test hypotheses is small. The present study may provide an additional demonstration of the value of cultural comparisons.

As background data let us note that reviews of data on handedness such as those by Wile (**1**) and Hildreth (**2**) show that while the frequency of left-handedness varies from activity to activity, in almost all studies the use of the left hand is more common among males than among females. The number of left-handed males have been found to exceed the number of left-handed females by 50% or to an even greater amount. This difference is present at least by four years of age and perhaps earlier.

The consistency of this finding suggests that this sex difference may have a biological basis. However, practically all of the investigations of hand preference have been conducted in Europe and America. For this reason the possibility exists that the lower frequency of left-handedness among women than among men is not biologically determined but rather that it may be a consequence of stronger social pressures against the use of the left hand among females than among males in western countries. In view of these rival interpretations of the data just reviewed, it seems worthwhile to report that in one Near Eastern country and probably in a much wider area the sex difference in handedness found in western countries does not exist.

The present study was conducted in the schools of Beirut, Lebanon, in 1955–56. Eleven schools were studied. In each classroom, a research assistant observed the handedness of children when they were engaged in writing in connection with their usual school work. In each school all grades from the kindergarten through Grade 5 were observed. Some of the schools were coeducational, others were not. In the case of noncoeducational schools, an attempt was made to match boys' schools and girls' schools with respect to socioeconomic class and religious affiliation.

A total of 2,656 pupils, 1,430 boys and 1,226 girls were observed. The frequency of left-handedness was found to be 5.0% among the boys and 4.9% among the girls. The small difference is statistically insignificant at the 5% level.

There is no reason to suppose that there are biological differences between Western and Near Eastern populations which would affect sex differences in handedness. The explanation of the difference between the finding just reported above and earlier findings probably is to be found in differences in social norms and child rearing practices. It seems likely that in the Near East left-handedness is no more reprehensible in women than it is in men. However, the precise attitudes and cultural conditioning in respect to handedness in this area can be identified only by further research.

[1] This study was conducted while the author was a visiting professor at the American University of Beirut. Expenses of the investigation were defrayed by a grant from the Rockefeller Brothers Fund. Adele Hamdan Taky Din and Leila Biksmati served as research assistants.

In further investigations it will be determined how widespread among Near-Eastern peoples is the sex equality in handedness ratios found in Lebanon.

This study suggests that it may be possible for a society to produce more sinistrality among females than among males. Whether such a society can be found remains to be determined.

REFERENCES

1. WILE, I. S. *Handedness: Right and left.* Boston: Lothrop, Lee and Shepard, 1934.

2. HILDRETH, Gertrude. The development and training of hand dominance. *J. genet. Psychol.*, 1949, **75**, 197–220, 221–254, 255–275, 1950, **76**, 39–100, 101–144.

Received April 4, 1958.

Journal of Educational Psychology
Vol. 49, No. 4, 1958

INSTRUCTOR EFFORT TO INFLUENCE: AN EXPERIMENTAL EVALUATION OF SIX APPROACHES[1]

E. PAUL TORRANCE

Bureau of Educational Research, University of Minnesota

AND

RAIGH MASON

L. G. Hanscom Air Force Base, Massachusetts

Instructors are frequently confronted with problems concerning the extent to which they should attempt to influence pupils in their attitudes and other behaviors having strong emotional overtones. They are uncertain about what techniques of influence are legitimate and the extent to which they should be persuasive. Some educational leaders are currently calling for teachers to be more persuasive in their efforts to influence pupil behavior. Others maintain that high pressure methods cause resistance or that such methods do violence to our democratic ideals.

Although most supervisors and instructors readily admit the importance of attitudes and other behaviors having strong emotional overtones, teachers generally have been reluctant to attempt to influence such behaviors. Especially have they shrunk from direct influence attempts. There has been a fairly pervasive attitude that a student's personal and social attitudes, emotional reactions, and the like are his personal business. This has been especially true in regard to matters affecting physical and mental health, eating, sleeping, sexual behavior, and the like. Seldom have such matters been chosen for scientific investigation and appropriate research situations have not been readily accessible.

Little or no scientific information exists to guide the decisions instructors and supervisors must make in deciding what kinds of effort to influence should be made. In social psychology, attention has been focused upon influence among group members (**5**), the influence of group norms (**1**), and the influence of associates in buying, politics, and the like (**4**). In the sales field, much has been said (**2**) about "low-pressure" selling in contrast to "high-pressure" selling. More recently, "no-pressure" selling appears to be coming into prominence (**2**). Little scientific research of an experimental nature has accompanied these trends, however.

One difficulty which has hampered research concerned with emotional reactions has been the unavailability of satisfactory criteria. Too frequently, it has been necessary to accept verbal expressions concerning such reactions. Even when it has been possible to obtain other behavioral measures, there has been doubt concerning the "real" emotional reaction behind the overt behavior. The authors have been fortunate in having access to a situation which provided a variety of criteria, including verbalized attitudes,

[1] This report is based on work done under ARDC Project No. 7723, Task No. 77461, in support of the research and development program of the Air Force Personnel and training Research Center, Lackland Air Force Base, Texas. Permission is granted for reproduction, translation, publication, use, and disposal in whole and in part by or for the United States Government. The opinions or conclusions expressed or implied herein are those of the authors. They need not be construed as necessarily reflecting the views or endorsement of the Department of the Air Force or of the Air Research and Development Command.

overt behavior, and an indicator of emotional response. The experimental situation involves the use of a survival ration commonly known as pemmican in the simulated survival exercise of the USAF Survival Training School. Use of the ration almost always elicits a wide range of response from extremely unfavorable to extremely favorable. Since the ration is recognized by most authorities in the field as the best available one for use in most survival situations, its use in training should increase its acceptability and, in fact, does (**9**).

An earlier study by the authors (**11**) gave a somewhat discouraging picture of the instructor's ability to influence the acceptability of the ration. When given scientifically developed information about the psychological, social, and training factors related to the ration's acceptability and asked to use this information on behalf of their crews, aircrew commanders (indigenous leaders) were far more successful than the crew instructors (**11**). Furthermore, those instructors who made the most effort to influence acceptability (as measured by statements made by both the instructor and the trainees) tended to obtain the lowest acceptability. Sustained efforts by indigenous leaders, however, were rewarded by increased acceptance.

Instructors in this and other situations frequently must face the very realistic problem of influencing the attitudes and emotionally toned behaviors of their students. Thus, it is evident that there is a need for a clearer understanding concerning what it is that instructors do which produces negative effects and what they can do to exert more positive influence. The purpose of the present study was to evaluate experimentally six alternative procedures by which training instructors may influence the acceptability of pemmican.

PROCEDURES

The *S*s of the study were 427 aircrewmen undergoing survival training. All *S*s received a double issue of the emergency ration consisting of a total of eight meat bars (pemmican) supplemented by chili and onion powder, two cereal bars, two fruitcake bars, 16 cubes of sugar, and eight packets each of soluble coffee and tea. During the nine-day simulated survival, escape, and evasion exercises, trainees were able to supplement these rations to some extent by such native foods as porcupine, crawfish, wild onions, water cress, camus, and the like.

A total of 43 instructors in the two successive classes were involved. Prior to the exercise, the training groups (crews consisting of 9 or 12 men each) were divided randomly into one control and six experimental groups. In each class, three training groups were involved in each of the experimental groups. In the first class, four groups were assigned to the control condition and in the second, three groups.

Instructors were briefed by three experienced psychologists thoroughly familiar with survival ration indoctrination and other aspects of the program of the USAF Survival Training School. The general purposes and design of the study were explained briefly. The instructors were asked to forgo their usual indoctrination procedures and use only the technique they would be assigned. Instructors then met in groups of three or four, as the case might be, with one of the experimenters to discuss the technique to which they had been assigned. Prior to the discussion, each instructor completed a questionnaire in which he indicated his personal reaction to the ration and described his usual indoctrination procedures. Each instructor was also given a typed sheet of instructions to be used as a guide in carrying out his assigned technique. The members of the control group were subjected to only the normal influences of the training situation.

The six experimental conditions may be described briefly as follows:

Experimental 1 (*No Influence*). Instructors were briefed to make no effort to influence trainees to accept the ration. They were instructed to say as little about it as possible, assuming a rather neutral stand on acceptability. They were cautioned, however, to avoid giving any impression of personal dislike for the ration. (This condition was designed to follow up a clue obtained from an exploratory study which indicated that trainees perceiving "no influence" attempts on the part of their instructors responded more favorably than those perceiving various degrees of efforts to influence.)

Experimental 2 (*Good Example*). Instructors were briefed to make no direct attempts to influence trainees to accept pemmican. They were issued a supply of the ration and instructed to make a definite attempt to manifest a definitely favorable attitude by personal example. This was done by eating the ration and casually expressing favorable reactions to it. They were cautioned to make no appeal to the trainees to eat the ration. (This condition was designed to evaluate the effectiveness of the often used admonition to instructors to "set a good example" and "never ask your students to do anything that you do not do.")

Experimental 3 (*Information*). Instructors were asked to give information about the value of the meat bar as an emergency ration and about ways of preparing it. They were instructed to give this information in an objective, factual, "take-it-or-leave-it" manner and to give no information about psychological reactions. (This condition was designed to evaluate what was considered a "low-pressure" technique of influence.)

Experimental 4 (*Group Explanation*). In addition to giving facts about values and ways of preparation, instructors were asked to emphasize the psychological factors which affect acceptability and to explain why this particular ration is used in the training exercise. The information about psychological influences was derived from previous research by the authors (**8, 9, 10**). (This condition was designed to test the value of using information gained through research to provide a psychological explanation of behavior.)

Experimental 5 (*Individual Explanation*). Instructors were briefed as in Experimental 4 except that they were asked to work with individuals as individuals instead of with the group. They were asked to appear natural and casual, but sincere, in their attempt to exercise personal influence.

Experimental 6 (*Evaluation*). Instructors were briefed to use the mildly coercive method of informing trainees that they would be "graded down" if they did not "really" try the ration. They were instructed to explain that failure to eat the ration was an indication of poor "will-to-survive," failure to take adaptive action, failure to take care of essential survival needs, failure to "play the game," etc. (This condition was designed to evaluate the effectiveness of using evaluation as a device for motivating or influencing behavior.)

Following the field exercise, all *S*s were administered a questionaire to obtain measures of acceptability and to provide additional facts concerning the conditions existing during the experiment. Acceptability items included: (*a*) the traditional hedonic scale (7-point), requiring the *S* to indicate his reactions to each of five methods of preparing pemmican; (*b*) the number of bars eaten; (*c*) reasons for not eating the remainder (made me sick, too greasy, smells bad, etc.); and (*d*) the conditions under which the *S* would use pemmican in the future. Previous research (**10**) had indicated that each of these items correlates significantly with and contributes importantly to an over-all index of rejection.

This over-all index of rejection was obtained by combining the items in the following manner. The ratings from the hedonic scales were weighted from one point for "like extremely" to seven points for "dislike extremely." If *S* indicated that he had not tried the bar according to one or more methods, the mean rating for the methods tried was assigned. One point was scored for each bar not eaten but no extra credit was awarded for eating more than the number of bars issued. Reports of having "been made sick" added five points and each of the other reasons for not eating the remainder of the bars was scored one point. Five points were added for responses of "would eat only when extremely hungry" and 10 points for "would not eat even if very hungry."

Results and Conclusions

First, an effort was made to determine the over-all effects of the six experimental influence techniques. Means and standard deviations for the Rejection Index and number of meat bars consumed and numbers and percentages for "made sick" and intension to eat the ration in the future "whenever hungry" for each condition are shown in Table 1. Using Bartlett's test, the requirements for homogeneity of variance are not met in the case of both the Rejection Index and number of meat bars consumed. Over-all chi squares indicate significant differences among the various conditions for both "made sick" and intension to eat the ration in the future "whenever hungry."

The heterogeneity of variance for Rejection Index and number of meat bars consumed seems to be due primarily to the small dispersion in Experimental 6. Table 2 presents the F ratios between Experimental 6 and each other condition. It will be noted that all of the F ratios are significant at the .05 level or better for Rejection Index. Only the F ratio between Experimental 2 and Experimental 6 fails to reach significance for number of meat bars consumed. In general, then, it may be concluded that all of the conditions are more erratic in their effects than Experimental 6.

Using the method described by Edwards (**3**, pp. 272–274) to correct for variance, direct tests were made to compare the means for Experimental 6 with the means for each other condition. The t ratios thus obtained are presented in Table 2. Using Rejection Index as the criterion, Experimental 6 appears to produce greater acceptability than the other conditions except Experimental 4 (Group Explanation). Using number of meat bars con-

TABLE 1

Means and Standard Deviations of Rejection Indexes and Number of Meat Bars Consumed and Percentage Made Sick and Intending to Use Bar in Future for Each Condition (Experimental and Control)

Condition	No.	Rej. Index		Bars Consumed		Made Sick		Eat in Fut.	
		Mean	SD[a]	Mean	SD[a]	No.	Pctg.[b]	No.	Pctg.[c]
Control	76	26.58	11.43	7.22	3.02	8	10.5	29	38.2
Exp. 1 (No infl.)	57	29.23	12.60	5.66	3.07	15	26.3	24	42.1
Exp. 2 (Good Ex.)	63	32.92	12.04	5.66	2.34	16	25.4	18	28.6
Exp. 3 (Info.)	62	27.73	12.93	7.95	4.60	11	17.7	38	61.3
Exp. 4 (Grp. Expl.)	61	25.61	11.54	6.75	2.94	17	27.9	32	52.5
Exp. 5 (Indiv. Expl.)	65	31.92	12.96	5.57	5.28	19	29.2	17	26.2
Exp. 6 (Evaluation)	43[d]	21.95	8.57	7.79	1.17	3	7.0	24	55.8

[a] Using Bartlett's test, requirements for homogeneity of variance not satisfied ($p < .001$)

[b] Chi square = 16.759; $df = 6$; $p < .02$

[c] Chi square = 27.227; $df = 6$; $p < .01$

[d] Number of Ss in Exp. 6 was reduced by eliminating one crew whose instructor was replaced by an unbriefe instructor after the beginning of the experiment.

TABLE 2

F Ratios and *t* Ratios Between Experimental 6 and Each Other Condition for Rejection Index and Number of Meat Bars Consumed

Condition	Rejection Index		Bars Consumed	
	F ratio	*t* ratio[b]	*F* ratio	*t* ratio[b]
Control	2.15[a]	2.48[a]	1.83[a]	1.14
Exp. 1 (No infl.)	1.96[a]	3.40[a]	1.93[a]	3.82[a]
Exp. 2 (Good Ex.)	2.26[a]	5.43[a]	1.12	4.61[a]
Exp. 3 (Info.)	1.80[a]	2.73[a]	4.33[a]	0.23
Exp. 4 (Grp. Expl.)	2.27[a]	1.84	1.77[a]	2.00[a]
Exp. 5 (Indiv. Expl.)	1.76[a]	4.77[a]	5.71[a]	2.71[a]

[a] Significant at the .05 level or better.

[b] Corrected for variance according to the method described by Edwards (3, Pp. 272–274).

sumed as the criterion, Experimental 6 achieved results significantly superior to all conditions except Experimental 3 (Information) and the Control Condition.

Further direct tests were made by comparing the Control Condition with each other condition. As already shown, Experimental 6 achieved significantly better results than the Control Condition using the Rejection Index as the criterion. Experimental 2 (Good Example) and Experimental 5 (Individual Explanation), however, appeared to produce significant negative effects (t ratios = 3.154 and 2.605 respectively, both significant at better than the .05 level).

When direct tests are made by applying chi-square analysis to the "made sick" criterion, the results are similar to those obtained for number of bars consumed. Experimental 6 again is superior at the .05 level or better to all conditions except the Controls and Experimental 3. Using intention of using the ration in the future as the criterion, Experimentals 3 and 4 produce results on a par with Experimental 6.

Several features concerning Experimental 3 (giving objective information about the value of the ration and describing methods of preparation) need to be noted. This experimental condition was accompanied by slightly higher mean consumption and willingness to eat the ration "whenever hungry" than even Experimental 6. The differences, however, fall far short of statistical significance. Using the latter criterion to compare Experimental 3 with the Controls, however, results obtained by Experimental 3 are superior (chi square = 7.31, significant at better than the .01 level). Experimental 3, however, appears to be quite erratic in its effects as indicated by the relatively large standard deviations of Rejection Index and number of meat bars consumed.

Discussion

In interpreting the results of this study, inescapable difficulties of experimental research in this area need to be made explicit. First, the *S*s under the control condition cannot be regarded as "untrained." They were subjected to varying degrees and kinds of influence. Questionnaire responses indicated that all of the control instructors conducted indoctrination concerning survival rations. It might even be argued that instructors in the experimental conditions were unpracticed and perhaps unskilled in the techniques which they were asked to use. It is certainly not contended that the instructors of the experimental groups were perfect in their adherence to the technique assigned. Nevertheless, the checks made indicated reasonable adherence to the assigned condition.

In general, the results of this study support the leads obtained from the previous studies to which reference has been made. It is interesting to note that the two methods having significant boomerang effects are those relying most heavily on personal influence. A number of explanations might be advanced. The explanation which best satisfies the authors is that the boomerang effect resulted from the phenomenon of "negative identification" discussed by Torrance and Ziller[2] in an earlier paper. According to this explanation, trainees perceive instructors as different from themselves and different in ways which prevent close identification. The trainee is a member of an aircrew. The instructor is not. He is an "earthbound" man. The instructor is something of a woodsman and is comfortable in the out-of-doors; usually, the trainee is not and frequently cannot imagine any "normal" person as being. The instructor is relatively young and in outstanding physical condition; usually the trainee is older and in comparatively poor physical condition. Thus, there appears to be the basis for negative identification and the adoption of behavior opposite that personally recommended by the instructor. These two techniques may also be regarded as "indirect" attempts to influence in comparison with the more "direct" approaches employed in Experimentals 3 and 6.

The experimenters' first impulse upon examining the results concerning the superiority of Experimental 6 was to reject them. Every attempt, of course, had been made in advance to maintain as rigorous controls as possible. The sampling, the indoctrination of instructors, and the collection of the criterion data had been accomplished as carefully as possible. The instructors did not see the completed blanks and *S*s were not required to sign their names, so there was little chance of threat to the trainee. Someone suggested, however, that trainees in Experimental 6 might have buried or destroyed some of their bars in order to make a good impression upon the instructor, since he was grading them on their use of it. Upon investigation, however, it was ascertained from independent witnesses that some of the crews in Experimental 6 had exhausted completely their supply of the ration and had bartered additional bars from other crews. For example, one crew bartered 33 additional bars and another 20 from other crews which did not consume their supply. The experiment was even replicated on another sample with essentially the same results.

Again, a number of alternative rationales might be advanced to explain the superiority of Experimental 6. Some might argue that men in our culture have been conditioned to respond favorably to this mildly coercive technique. If this were the only explanation, however, one would expect more evidence of "behavior without conviction" than is apparent. In the light of previous studies, the authors would argue that this is a simple and direct technique which is superior to indirect types of influence. Survival ration indoctrination is made an integral part of training and given its proper importance. The ration takes on meaning in terms of training and preparation for possible future emergencies and/or extreme conditions. It is no longer "just something to eat during training." This type of influence attempt places the instructor in an "official" rather than a "personal" role and he is probably more acceptable and influential in such a role.

It should not be concluded that instructors should avoid the "good example" and other techniques of personal influence. According to our interpretation, however, such techniques are likely to boomerang if trainees identify negatively with the instructor. Even Experimental 4 (Group

[2] Torrance, E. P., & Ziller, R. C. *Negative identification in groups as a function of personality differences.* Reno, Nevada: Survival Methods Branch, Air Force Personnel and Training Research Center, Stead Air Force Base, March 1956. (*Laboratory Note* CRL-LN-210.)

Explanation) is probably influenced by this phenomenon. The rationale for Experimentals 4 and 5 was taken from Stefansson's experiences in indoctrinating members of his exploration parties concerning "arctic hysteria" (7). He maintained that newcomers to the Arctic were simply not bothered by "arctic hysteria," if they were given a satisfactory explanation of its psychological basis. It is likely that these young explorers identified strongly with Stefansson, accepted his explanation and were influenced by it.

The findings are probably applicable to situations in which instructors need to influence attitudes and other behaviors with strong emotional overtones. In general, it would appear that instructor attempts to influence should be of the direct, "take-it-or-leave-it" variety and should be made in the instructor's "official" rather than "personal" role. Although the influence of associates may be far stronger than that of instructors, the findings of this study do suggest that instructors may play significant roles in influencing attitudes and other behaviors having strong emotional overtones and that this can be a fruitful area of research. It is possible that the findings of this study can be generalized to other conceptually similar situations where it is desirable to influence attitudes and behavior, particularly in educational situations. It is also likely that some of the findings may apply to influence situations in such activities as selling. The findings are quite in accord with theories which have been developed in the past decade concerning the superiority of "low-pressure" sales techniques. Naturally, all of these findings need to be tested in other situations.

SUMMARY

A sample of 427 aircrewmen participating in a survival exercise were divided randomly into seven groups (six experimentals and one control). Crew instructors of the experimental crews were requested to conduct the survival-ration indoctrination according to specific instructions. Using four criteria of acceptance of the ration, an experimental condition making the food indoctrination a regular part of the training accompanied by evaluation tended to produce superior results. Promising results were also obtained from a "low-pressure" technique relying chiefly upon objective information and straight-forward instructions concerning preparation. Significant negative effects were obtained from conditions relying upon personal persuasiveness, setting an example, and the like.

REFERENCES

1. BERKOWITZ, L. Group norms among bomber crews: Patterns of perceived crew attitudes, "actual" crew attitudes, and crew liking related to aircrew effectiveness in Far Eastern combat. *Sociometry*, 1956, **19**, 141–153.
2. BURSK, E. C. Thinking ahead: Drift to no-pressure selling. *Harvard Bus. Rev.*, 1956, **34**, 25–32f.
3. EDWARDS, A. L. *Statistical methods for behavioral sciences.* New York: Rinehart, 1954.
4. KATZ, E. & LAZARFELD, P. F. *Personal influence.* Glencoe, Ill.: Free Press, 1955.
5. LINDSEY, G. (Ed.) *Handbook of social psychology.* Cambridge: Addison-Wesley, 1954.
6. MCNEMAR, Q. *Psychological statistics. (2nd ed.)* New York: John Wiley, 1954.
7. STEFANSSON, V. *Arctic manual.* New York: Macmillan, 1953.
8. TORRANCE, E. P. Training factors affecting survival ration acceptability. In *Conference Notes, Food Research and Development Coordination Conference, Wright-Patterson Air Force Base, Ohio, 9–10 October 1956.* Chicago: Quartermaster Food and Container Institute for the Armed Forces, 1957. Pp. 74–90.

9. Torrance, E. P. Sensitization versus adaptation in preparation for emergencies: Prior experience with an emergency ration and its acceptability in a survival situation. *J. appl. Psychol.*, 1958, **42**, 63–67.

10. Torrance, E. P., & Mason, R. Psychological and sociological aspects of survival ration acceptability. *J. clin. Nutr.*, 1957, **5**, 176–179.

11. Torrance, E. P., & Mason, R. The indigenous leader in changing attitudes and behavior. *Int. J. Sociometry*, 1956, **1**, 23–28.

Received June 2, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 4, 1958

OVERLAP AMONG DESIRABLE AND UNDESIRABLE CHARACTERISTICS IN GIFTED CHILDREN

GORDON LIDDLE

University of Chicago

Terman's study of the gifted has shown that, in general, highly intelligent children in addition to being larger and healthier, are also somewhat more adjusted socially than the average child. His gifted group, including those who had been accelerated in school, carried these advantages into adult life (**3**). More recently, the Ford Foundation's Fund for the Advancement of Education found that a group of gifted children coming to college two years early had adjusted as well socially and emotionally to college life as had their classmates (**2**).

At present there is considerable discussion in educational circles of the advantages and disadvantages of acceleration and special grouping as administrative tools in meeting the needs of gifted children. Many school administrators look with disfavor on these techniques. When asked why, they usually give a reply which implies that the social adjustment of gifted children is rather fragile. Is this fear justified? Are gifted children more often or less often subject to severe maladjustment than other children?

The purpose of this study is to examine the overlapping of talents and maladjustments in a group of 1015 public school children in late childhood and early adolescence. The research is part of a 10-year action-research project being carried out by the Committee on Human Development of the University of Chicago.

PROCEDURES

The population of the study comprised the entire public school population of the fourth and sixth grades in a Midwestern city of 45,000 in the school year 1951–52, the first year of the study. For each child included in the population, the following characteristics were measured: aggressive maladjustment, withdrawn maladjustment, social leadership ability, artistic talent, and intellectual ability. Tests designed to measure all these characteristics were administered during the first year of the study. The tests measuring the first three characteristics were readministered during the second and fourth years of the study. Children for whom test information was incomplete were excluded from this study.

Two tests were used in determining aggressive maladjustment, withdrawn maladjustment, and social leadership ability. One is the "Who Are They?" (W.A.T.), a sociometric instrument based on children's evaluations of their peers with respect to these three behavioral characteristics (**1**). A child's leadership score was determined in response to questions such as, "Who are the leaders, the leaders in several things?" "Of the people you run around with, who are the ones who come up with good ideas of interesting things to do?" Aggressiveness was determined by nominations to questions such as, "Who are the boys and girls that seem to be against everything that is suggested—the gripers?" "Who are the bullies, the boys and girls who try to push others around?" The following questions are typical of those contributing to the withdrawn score, "Who are the ones that are too shy to make friends easily? It is hard to get to know them." "Who are the boys and girls who usually come and go alone and stay by themselves most of the time, even though they aren't trouble makers?"

The other instrument used to measure aggressiveness, withdrawnness, and leadership was the "Behavior Description Chart" (B.D.C.), a forced-choice teacher rating instrument. Here teachers had to pick the items "most like" and "least like" a given child in a series of 10 groups of five statements each, such as the following:

A. Other people find it hard to get along with him.

B. Is easily confused.

C. Other people are eager to be near him or on his side.

D. Is usually willing to go along with the group.

E. Interested in other people's opinion and activities.

In the foregoing pentad, if A was thought to be the statement "most like" this child, this contributed to his aggressive score. If B was thought to be most typical, this contributed to his withdrawn score. Item C is a leadership item, and D and E are not scored since they are presumed to be typical of average children. Similarly a "least like" nomination for A, B, or C subtracted from the child's score on that variable.

Each individual was given a percentile score for aggressiveness, withdrawnness, and social leadership ability on each of the two tests administered in each of the three years. Because high scores for one year might be unduly affected by a temporary upset in the child's life or an atypical relationship with one of his teachers, it was thought best to add the six percentile scores obtained from the two tests and divide by six to get a local mean percentile score. This was done for each of the three behavioral characteristics.

TABLE 1

Correlations of Percentile Scores from Two Year's Tests

Characteristics	Tests	
	Who Are They?	Behavior Description Chart
Aggressive maladjustment	.40	.54
Withdrawn maladjustment	.47	.63
Social leadership ability	.74	.63

The scores from the first two years of testing have been utilized for a reliability study. Product-moment correlations between the two sets of percentile scores are reported in Table 1.

It should be remembered that this is a severe reliability measure, since in the second year the children were in different classrooms, with different teachers, and with from 25 to 60 per cent turnover in classroom membership.

It was noticed that the children who ranked high in any one category generally ranked in the same category on subsequent tests, but that considerable shifting occurs in the relative positions of the low-ranking children. Measured leadership ability remained more constant from one year to the next than did the maladjustment characteristics.

For all three characteristics, the top 7–10 per cent of the children received half of all the nominations on the W.A.T. Thus, this instrument differentiates quite clearly among those children displaying each characteristic to a high degree, but does not differentiate among those who seldom display the characteristic being measured. The B.D.C. yielded a rather similar distribution of scores.

Intellectual talent was determined through use of both tests of "general" intelligence and tests of such "specific mental abilities" as could be measured in children of 10 or 12 years of age. Also an effort was made to include some tests which were thought to be more "culture-fair"; that is, tests which did not discriminate against the children of lower socioeconomic status groups.

The following tests were used for each child: the Science Research Associates

Primary Abilities Test (P.M.A.) for ages 7–11, the Davis-Eells Games, the Goodenough Draw-A-Man Test, the Thurstone Concealed Figures Test, and the verbal, spatial, and reasoning subtests of the Chicago P.M.A. for ages 11–17.

The percentile scores on the seven intellectual measures were averaged. This was a rather arbitrary decision, but it might be said that the use of a multiple-regression equation was discarded since this method requires an accepted, independent criterion of talent with which the screening instruments could be correlated. Academic achievement test scores or academic grades could have been used as criteria, but there was little reason to suppose that they would have been better than the test score itself as a criterion.

Artistic talent was determined by asking a group of local artists to rate four pictures drawn by each child. These pictures were: a classroom as seen from the doorway, a landscape, a free assignment to draw the child's favorite subject, and the Goodenough Draw-A-Man Test scored with different criteria from those used to score it as an intelligence test.

After the testing had been completed, the 10% of the total group displaying each of the five characteristics to the highest degree were set aside, and it is these top 10% groups which will be investigated in this study.

Results

Table 2 points out that children who are talented in one area are quite likely to be talented in other areas, but are quite unlikely to be seen as highly maladjusted. Chi square was used in determining the statistical significance of the differences between observed and expected frequencies of overlapping among the five characteristics.

Table 2 shows that:

1. Social leadership ability is positively related to the other talents and negatively

TABLE 2

Overlapping of Talent and Maladjustment Categories

(1015 children)

Characteristic	Leadership Ability ($N = 104$)	Intellectual Ability ($N = 107$)	Artistic Talent ($N = 102$)	Withdrawn-ness ($N = 103$)	Aggressive-ness ($N = 101$)
Leadership	—				
Intellectual		—			
Observed	45.00*				
Expected	10.96				
χ^2	131.66				
Art			—		
Observed	31.00*	33.00*			
Expected	10.45	10.75			
χ^2	50.05	57.23			
Withdrawn				—	
Observed	.00*	3.00*	5.00		
Expected	10.55	10.86	10.35		
χ^2	13.08	7.08	3.42		
Aggressive					—
Observed	.00*	3.00*	7.00	8.00	
Expected	10.35	10.65	10.15	10.25	
χ^2	12.81	6.76	1.21	.61	

* 1% level of confidence

related to the two maladjustment characteristics. There are 76 instances of overlapping between social leadership and the other talent areas, but no overlapping with the two maladjustment areas.

2. Intellectual talent is significantly related to the other talents and is almost as surely negatively related to the maladjustment characteristics. There are 78 instances of overlapping with the other two talent areas, but only 6 instances of overlapping with the two maladjustment characteristics. Intellectual talent and social leadership ability overlapped more than four times as often as would be expected on the basis of chance occurrence.

3. Artistic talent is highly related to the other talent areas, but while there is a negative relationship between artistic talent and the maladjustment characteristics, this relationship is not statistically significant. There are 64 instances of overlapping with one of the talent areas, while there are 12 instances of overlapping with one of the maladjustment categories.

4. The overlapping between withdrawn and aggressive maladjustment is not statistically significant.

Since there is a possibility that only the extremely intellectually gifted have severe adjustment problems, it was decided to examine the 51 children with the highest intellectual scores, the top 5%. Only two of the 51 children were in the top 10% in one of the maladjustment categories, while there were 41 instances of overlapping with the talent areas. Thus, the top 5% in intellectual talent had more overlapping with the other talents and less overlapping with the maladjustment categories that did the second 5%.

TABLE 3

Intercorrelations of Talent and Maladjustment Categories

Variables	Intellectual	Leadership	Withdrawn	Aggressive
Intellectual	—	.49*	−.45*	.05
Leadership	.37*	—	−.76*	−.05
Withdrawn	−.28*	−.61*	—	−.22*
Aggressive	−.11	−.23*	−.24*	—

Note.—Coefficients for girls ($N = 143$) above diagonal; for boys ($N = 130$), below diagonal.

* 1% level of confidence

For readers who are interested in the correlations of these characteristics throughout their entire range, Table 3 presents these correlations for 273 of the children who were in the sixth grade at the beginning of the study. It is believed that the correlations for the entire 1015 children would be quite similar.

In interpreting Table 3 it must be remembered that the tests used to measure leadership and the maladjustment characteristics were set up to screen out those children displaying a given characteristic to a high degree and were not intended to differentiate between children displaying these characteristics to a lesser degree.

The table indicates that intellectual ability and social leadership ability are significantly correlated and that both are negatively related to withdrawnness. While the negative relationship between withdrawnness and aggressiveness is statistically significant, it is not extremely high. Except for the negative relationship between aggressiveness and leadership for boys, there are no statistically significant relationships between aggressiveness and the talent variables. Artistic talent was not quantified throughout the entire range and thus could not be correlated with the other variables.

Summary

The top 10% groups in intellectual talent, social leadership ability, artistic talent, aggressive maladjustment, and withdrawn maladjustment were examined. It was found that children who were highly gifted in one of the three talent areas were quite likely to be talented in other areas, and quite unlikely to be seen as highly maladjusted by their teachers and classmates.

REFERENCES

1. Bowman, P. H., et al. Mobilizing community resources for youth. *Supplementary educational monographs*, 1956, No. 85, Chicago: University of Chicago Press.

2. Ford Foundation, Fund for the Advancement of Education. They went to college early. *Evaluation report* No. 2, New York: Author, 1957.

3. Terman, L. M., & Oden, Melita H. The *Gifted child grows up*. Standford: Stanford Univer. Press, 1947.

Received June 22, 1957.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 4, 1958

ATTITUDE CHANGE THROUGH UNDIRECTED GROUP DISCUSSION

K. M. MILLER AND J. B. BIGGS[1]

University of Tasmania

Studies of attitude change have emphasized many variables; the present study was concerned with the effectiveness of free group discussion about racial groups when the discussion groups are sociometrically structured. Some writers, (**3, 8, 11**) have stressed the importance of using sociometric structure as an aid to effective classroom work, suggesting that the educational process is more efficient when groups are composed of mutually attracted members, i.e. when groups are cohesive.

DESIGN OF STUDY

Two types of groups were selected from a class—psychegroups considered high, and sociogroups low in cohesion. Attitude change was assessed by testing with an attitude scale before and after a period of undirected discussion about a number of racial groups. The stability of any change was assessed by a third test some weeks later. A control class completed the attitude scale at the same times but without intervening discussion.

The interval between the first and second tests was designed to minimize memory of responses to the first. School vacation prevented the interval between the second and third tests being identical with the first interval. On no occasion were *S*s informed that subsequent tests would be given.

TECHNIQUES

Sociometric. The conventional form of the Moreno technique was used, *S*s being asked to write the names (up to five for each category) of those classmates next to whom they would like to sit and would not like to sit.

Social Attitudes. A Bogardus type scale was selected as the most suitable both for repeated measurement and for showing changes after discussion. The form used was similar to the Zeligs and Hendrickson (**13**) modification but had been independently derived by the senior author for a previous study. The steps were:

I would like to have live in my home.
I would like to have as a close friend.
I would like to go for a holiday with.
I would like to have in my sports team.
I would like to work with in school.
I would like to have live in my street.
I would like to have live in my country.

So the list of racial groups would be meaningful for the *S*s, 14 were selected either on the basis of percentage of national group among migrants to Australia or for historical reasons, e.g., English and Japanese. The groups were: American, Chinese, Dutch, English, German, Indian, Irish, Italian, Japanese, Jewish, Negro, Polish, Russian, Balt.

Scoring was by the Zeligs method (**12**) whereby each positive response counted one point.

SUBJECTS

Two third year secondary school classes of 26 and 16 boys respectively were selected. Of the larger class 24 members with a mean age of 177 months, *SD* 8.5, were experimental *S*s while all of the smaller class with a mean age of 180 months, *SD* 9.0 months, were control *S*s. The difference in age was not significant.

SELECTION OF DISCUSSION GROUPS

The method of analysis of sociometric data suggested by Clark and Maguire (**2**)

[1] Now at National Foundation for Educational Research, London.

was used, supplemented in the final choice of groups by reference to sociograms. Three friendship or psychegroups and three neutral or sociogroups each containing four boys were selected. The psychegroups were chosen so that within each (i) all members expressed strong 1, 2, 3) choices for each member of the group (ii) each member received at least one mutual choice and (iii) no one was rejected by other members of the group. The composition of the sociogroups was such that no member had expressed either acceptance or rejection of any other member.

Procedure

The task was presented as part of a general study being conducted in several countries, to find out how children thought about people in their own and other countries.[2] To encourage frankness *Ss* were assured that all replies were confidential and would not be seen by anyone in the school. The sociometric and social distance scales were then administered to both control and experimental classes.

The initial administration of the social distance scale was as recommended by Bogardus (**1**), the *E* reading the items at three-second intervals. On the later occasions *Ss* were allowed to complete the scale at their own rate.

Four weeks later the group discussions were begun, friendship and neutral groups working alternately. Each group was instructed only after assembling and at the conclusion of the discussion the members were asked not to discuss the activity with the other boys. The *E* introduced the discussion, explaining that he would like each member to say something about a number of racial groups. The discussion period lasted approximately 30 minutes, two minutes being allowed for each of the 14 races, though discussion of any one group was not abruptly terminated.

During the discussion, *E* was passive and nondirective, averting any questions put directly to him. Following the discussion the social distance scale was readministered. After half of the group discussions had been completed the scale was given a second time to the control group.

As a check on the stability of any change the scale was given a third time two weeks after the last group discussion to both control and experimental classes.

Results and Analysis

Comparison of control and experimental groups. A t test of initial scores revealed no significant difference ($t = 1.10$, 38 df), thus showing that the mean attitude level of the two could be considered equivalent.

The effect of discussion. The mean scores for each of the three sections—psychegroups, sociogroups, and control class—on first and second administrations were tested for differences. The differences for the friendly and neutral *Ss* were significant beyond the one per cent level ($t =$ 3.43 and 3.16, 11 df respectively) while the difference for the controls was not significant, ($t = 0.32$, 15 df).

Differences between second and third administrations. Inspection of Table 1

TABLE 1
Mean Social Distance Scores for Friend and Neutral *Ss* by Group

	Administration		
	1	2	3
Friends			
Group 1	33.25	47.25	42.25
Group 2	49.25	72.50	57.50
Group 3	52.00	54.25	55.75
Total	43.75	58.00	51.83
Neutrals			
Group 1	39.50	51.25	50.50
Group 2	37.75	70.00	63.00
Group 3	51.00	57.75	54.30
Total	42.75	59.67	56.00

[2] A report is in preparation on the International study, the design of which is described in (**7**).

TABLE 2
NUMBER OF REMARKS OF EACH TYPE MADE DURING DISCUSSION

	Favourable	Unfavourable	Neutral	Total
Psychegroups	151	133	42	326
Sociogroups	155	129	39	323

shows that the mean scores for the third administration are lower than immediately after the discussion period in the case of the friendly and neutral *S*s. These differences were not significant for any of the three sections (*t* values of 2.14, 1.27, and 0.86 for friends, neutrals, and controls respectively).

Differences between first and third administrations. A similar analysis was made of differences between first and third sets of scores. The *t* test analysis showed that final scores were significantly greater than the initial scores for both psychegroup and sociogroup *S*s; at the two per cent level for the former and at the five per cent level for the latter. Again the differences in the control class scores were not significant ($t = 0.33$).

Quantitative aspects of the discussions. The remarks made by each *S* were recorded and categorized as favorable, unfavorable or neutral towards the race under discussion. The total remarks are shown in Table 2 where it is seen that the total number of remarks and the distribution according to category are approximately equal for the psychegroups and sociogroups. Examination showed that in the psychegroups remarks were somewhat more evenly spread over all members than in the sociogroups.

DISCUSSION

The analysis has shown that the members of both psychegroups and sociogroups show more tolerance (decreased social distance) after free undirected discussion of some characteristics of different racial groups. No such change is shown by control *S*s tested after the same interval of time. The amount of initial change appears to be unrelated to type of group as the mean difference between friends and neutrals is not significant.

When the same scale is applied again after an interval of two weeks the mean scores of members of both types of experimental group are closer to the mean scores in the first testing than they were on the second. The amount of this reversion is, however, not statistically significant, indicating that the favorable change engendered by the discussion was fairly stable over this short period. Further confirmation was provided by the comparison of the initial and final scores which were significantly different for both psychegroups and sociogroup subjects, at two and five per cent levels respectively.

Few, if any, investigators have suggested that free undirected discussion about racial groups would lead to the measurable changes demonstrated in this study. Some investigators (**6, 9**) indicate that attempts to change attitudes are more successful when *S*s are members of naturally working groups. While a class is in some respects a functioning group it has within it a number of groups which are more cohesive than the class as a whole. Thus it would not have been surprising to find the members of the psychegroups showing greater and more stable change than the sociogroup members. The results of the present study are not in accord with such expectations as both friends and neutrals show (approximately) equally significant changes and stability of change. Moreover, they did not differ from each other in degree of change throughout.[3]

The present findings are, however, in keeping with the suggestion that close

[3] Unrelated *t* tests on differences scores between psychegroup and sociogroup members at each stage were nonsignificant.

friendships may lead to better communication and wider participation in the group (**6**). The evidence for better communication is provided by notes of the actual discussion sessions. The climate in the psychegroups was freer and more lively and discussion more spontaneous than that in the sociogroups where discussion tended to be more formal and reserved, though no less frequent, and not different in the proportion of favorable, unfavorable and neutral remarks. Thus although quantitative differences between friend and neutral groups were not discovered, there is a difference in the way in which the quantitative result was achieved.

Some investigators (**8, 10**) have considered such changes in terms of group conformity, suggesting that there would be a greater tendency for members of sociogroups, being less cohesive than psychegroups, to establish some norm or common position. Others (**4, 9**) have considered such changes as a function of security—insecurity and personality adjustment—maladjustment suggesting that less secure, less well-adjusted persons may, as a means of establishing a more secure group relationship, change towards a central position. When the results of the present study were examined for evidence of conformity it was found that for all three sociogroups and for none of the psychegroups the range of scores after discussion was considerably smaller than before.

The relevance of these findings for education requires consideration as the results seem to be at variance with those usually claimed by proponents of the sociometric approach. Work of investigators such as Cunningham, Oeser, and Shoobs suggests that group discussion would be more effective in the psychegroups than in the sociogroups, whereas it has been shown in this study that measured changes are approximately equal for both types of groups. Further research is required to ascertain whether the changes are primarily a function of the learning process, in one case, and a function of personality factors and insecurity in the other.

Summary and Conclusions

This study was an attempt to relate attitude change with free discussion in psyche- and sociogroups. Several findings are definite while others merit further investigation.

1. Free, undirected discussion about racial groups by two types of small groups, selected on a sociometric basis, resulted in a significant change of attitude irrespective of the type of group. Further, this change was relatively stable over a short period.

2. Contrary to expectations from sociometric studies in the classroom, and from studies of group structure, no significant differences between the quantitative changes of friendly and neutral *S*s were discovered.

3. Nevertheless, it was suggested that the psychological processes in the two types of groups might be different and that further investigation is necessary to show whether the tendency for the scores of members of sociogroups to come closer to a central position after discussion is a function of personality adjustment.

REFERENCES

1. Bogardus, E. S. Measurement of personal group relations, *Sociometry*, 1947, **10**, 306–311.
2. Clark, R. A., & McGuire, G. Sociographic analysis of sociometric valuations *Child Develop.*, 1952, **23**, 141–154.
3. Cunningham, Ruth, et al. *Understanding group behavior of boys and girls*, New York: Teachers College, Columbia Univer., 1951.
4. Grossman, B., & Wrighter, J. Relationship between selection-rejection and intelligence, social status, and per-

sonality amongst 6th grade children. *Sociometry*, 1948, **11**, 346–355.

5. Kelley, M., & Thibaut, J. Experimental studies in group problem solving and process, in G. Lindzey (Ed.), *Handbook of social psychology*. Cambridge, Mass., Addison Wesley, 1954, p. 762.
6. Lewin, K., & Grabbe, P. Conduct, Knowledge, and Acceptance of New Values, *J. Soc. Issues*, 1945, **1**, 53–74.
7. Miller, K. M. Evaluation in adult education. *Internat. soc. Sci. Bull.*, **7**, 1955, 430–442.
8. Oeser, O. A. *Teacher, pupil and task*. London: Tavistock, 1955.
9. Sherif, M., & Sherif, C. W. *Groups in harmony and tension*. New York: Harper, 1953.
10. Sherif, M., & Sherif, C. W. *Outline of social psychology* (Rev. ed.) New York: Harper, 1956, pp. 301–328.
11. Shoobs, N. E. Sociometry in the classroom, *Sociometry*, 1947, **10**, pp. 233–241.
12. Zeligs, Rose. Children's intergroup attitudes. *J. Genet. Psychol.* 1948, **72**, pp. 101–110.
13. Zeligs, R., & Hendrickson. Racial attitudes of 200 6th grade children. *Sociol. soc. Res.*, **18**, pp. 26–36.

Received July 26, 1957.

THE JOURNAL OF
EDUCATIONAL PSYCHOLOGY

Volume 49 October, 1958 Number 5

RELATIONSHIP OF SELF-ACCEPTANCE TO OTHER VARIABLES WITH SIXTH GRADE CHILDREN ORIENTED IN SELF-UNDERSTANDING[1]

PAUL BRUCE[2]

Child Welfare Research Station, State University of Iowa

The purpose of this study was to investigate with sixth grade children some of the relationships between a measure of self-acceptance and other personality variables. The relationships were studied in two groups of *S*s: one consisting of pupils who had taken part in a learning program designed to develop a more understanding and analytical approach to their own and others' behavior (the experimental group) and one group consisting of pupils who had not undertaken such a program (the control group).

A review of the literature—particularly the writings of Rogers (**15**) and other researchers with a client-centered therapy orientation—suggested that self-acceptance be defined as the congruence between the way a person thinks himself to be (self-concept) and the way he thinks he would most like to be (ideal-self). Accordingly, in this study, a measuring instrument from which Self-Ideal Discrepancy scores were obtained was devised suitable to a sixth grade population.

Within the framework of a self-ideal discrepancy measure of self-acceptance, a child's indication of his ideal-self can be viewed as his expression of how he feels some of his psychological needs (such as the needs for security, personal worth, status, etc.) can be best satisfied. In other words, underlying the conception of ideal-self is the assumption by the child that if he becomes more like his ideal-self, he will be better able to satisfy some of his secondary needs. For example, implicit in a child's wanting "to be better looking," "to have more friends," "to be less sensitive," etc., is the feeling that one's affectional, security, and status needs (among others) would be better satisfied if these ideals were achieved. If a child, then, feels he is quite unlike his ideals or feels he is not making progress towards his ideals, it might be speculated that adequate satisfaction of several of his secondary needs is being blocked in some way, and thus several predictions concerning his behavior might be made.

It would be expected that a child who indicates a marked discrepancy between his self-concept and ideal-self, feeling his need satisfaction blocked, would show evidences of insecurity in his behavior and would yield responses indicating manifest anxiety. Such relationships between measures of self-acceptance and various measures of insecurity and anxiety have been obtained with subjects of college age and beyond (**1, 2, 7, 13, 16**). One of the major purposes of this investigation was

[1] This paper is based on a doctoral dissertation submitted to the Graduate College of the State University of Iowa in 1957. The writer is indebted to Ralph H. Ojemann for his encouragement and guidance throughout the study and to the members of the Preventive Psychiatry Project staff for their invaluable assistance.

[2] Now at San Diego State College.

to see if the relationship between self-ideal discrepancy (as a measure of self-acceptance) and measures of observed insecurity and manifest anxiety held for a group of sixth grade children.

The orientation program which the experimental group of *S*s had undertaken was designed to help each pupil develop a "causal," analytical orientation towards his social environment. The program was based on the assumption that a child who is provided with the opportunity for understanding some of the many causes underlying his own behavior and the behavior of people about him will be able to make more effective adjustments. Involved in the program was the special training of the teachers and the use of certain special curricular materials, the details of which have been published elsewhere (**9, 11, 12**).

With respect to this investigation, the orientation program provided a group of *S*s who had been trained in self-understanding, among other things. Several writers, such as Jersild (**6**) and Rogers (**14**), have indicated that an important characteristic of a self-accepting person is self-understanding. The question arises, then, as to whether a child with increased self-understanding is thereby more self-accepting. It might be hypothesized that as a child gains insight into his own motivations and dynamics—that is, gains in self-understanding—he would give evidence of being more self-accepting. Also to the extent that the orientation program helps the individual learn how to work out his daily situations more effectively, he may be helped to make progress towards his ideal thereby reducing the self-ideal discrepancies.

On the other hand, increased self-understanding may not affect self-acceptance, particularly when self-acceptance is defined in terms of self-ideal discrepancy. For example, an individual may see himself some distance from his ideal, but with increased self-understanding, understand some of the reasons for this situation. In such a case, self-ideal discrepancy might remain high even after the orientation program, although the individual would feel better—that is, be less anxious—about it. Thus, the orientation program might serve to reduce anxieties which normally accompany high self-ideal discrepancy without necessarily reducing the discrepancy itself. It appears, therefore, that the orientation program could have diverse effects upon a self-ideal discrepancy measure of self-acceptance.

The specific questions to which answers were sought in this study are: (*a*) Is there a significant difference between *S*s with high self-ideal discrepancy and those with low self-ideal discrepancy in average scores on a test of manifest anxiety? (*b*) Does this relationship hold for *S*s who have taken part in an orientation program in self-understanding? (*c*) Is there a significant difference between *S*s with high self-ideal discrepancy and those with low self-ideal discrepancy in average ratings on an observation scale of insecurity?

Procedure

Measures used in study. A Self-Acceptance Scale was devised and initially administered to several sixth grade classes. The results and items were analyzed qualitatively and quantitatively. The 10 items finally selected for the revised Self-Acceptance Scale, in general, reflected affective characteristics about which individuals in our culture were thought to have substantial feelings. (Only nine of these items were scored since one proved ambiguous in subsequent administrations of the scale.) A more detailed description and a copy of this scale appears elsewhere (**3**).

Examples of typical items are these: "1. This is someone who feels that others don't like him or her—someone whom nobody seems to care about;" and "8.

This person is happy and cheerful most of the time—one who seems to enjoy what he or she does." Some of the items in the scale were adapted from the Social Analysis of the Classroom Inventory developed by the Horace Mann-Lincoln Institute of School Experimentation (**5**). Descriptive statements were utilized in place of the trait-names (used on most other scales of this type) in order to reduce the amount of ambiguity.

The *S*s were asked to answer four questions about each of the descriptive statements:

1. Would most boys and girls my age like to be like this person? (This question was not scored but was included as a check on the general favorability or unfavorability of each item.)
2. Am I like this person? (Self-Concept)
3. Am I becoming more like this person? (The results obtained by the use of this question were inconclusive and are not discussed in this paper.)
4. Would I like to be like this person? (Ideal-Self)

The *S* answered each question indicating to what extent he felt it was true by checking along a 5-point scale reading: "very much so," "quite a bit," "somewhat," "not very much," and "not at all."

The particular score yielded by this scale which was used in the major analyses reported in this paper is the Self-Ideal Discrepancy score. This score is the discrepancy between the rating in answer to Question 2 (Self-Concept) and the rating in answer to Question 4 (Ideal-Self) summed for the nine items. Test-retest reliability for this score is reported in the next section.

The other measures used in this study included the Children's Manifest Anxiety Scale developed by Castaneda et al. (**4**) and the Kooker Security-Insecurity Rating Scale. The Children's Manifest Anxiety Scale consists of 42 anxiety items which are answered "yes" or "no." The Kooker scale (**8**) gives evidence of children's behavior which is independent of their responses on a paper-pencil type test and which reflects their "typical" behavior patterns. This scale requires a trained observer who is familiar with the child's behavior to rate him on a series of 19 behavior items, as occurring "frequently," "fairly often," or "seldom." In the present study, each classroom was visited for a five-day period and the children rated by one of the two trained observers (who remained ignorant of the nature of this investigation). Both observers practiced rating the same *S*s with the Kooker scale until they were in substantial agreement with each other.

With the exception of the Kooker scale, the other measures were administered to all classes by the investigator. The teachers were not involved in the measurement aspects of the investigtion and were not aware of the nature of the study.

Subjects of investigation. The subjects in this study were pupils of eight sixth grade classes in different elementary schools located in comparable neighborhoods with respect to socioeconomic status. The four experimental classes had undergone the orientation program described above. Most of the pupils in two of these experimental classes had been in the program for two consecutive years; whereas the pupils in the remaining two experimental classes had undergone the orientation program for the current year only.

At the beginning of the program, four control classes were selected controlling for some of the teacher variability. Thus, for every teacher of an experimental class, a control teacher was selected and matched on the basis of several variables including age, sex, number of years teaching experience, and educational level. Unlike the experimental teachers, the control teachers did not have special training for this program and did not carry on the planned learning program in their classes.

With respect to the *S*s themselves, in the experimental classes 50 boys and 48 girls completed all of the testing involved in this study, while in the control classes tests were completed by 40 boys and 46 girls. The average IQ for the experimental classes was 105.9 and 105.4 for the control classes as measured by the Otis Self-Administering Test of Mental Ability. As far as could be determined from school officials, no systematic method was used to assign the pupils to particular classes or to teachers (except for the two-year experimental group who were kept together the second year specifically for the orientation program.) Thus, with the *S*s coming from comparable home backgrounds, being randomly placed in their particular class groups, and being undifferentiated with respect to average intelligence scores, the pupils in all eight classes were assumed to be originally a random sampling of a population of such pupils in their particular schools.

Statistical analyses. To test two major problems of this investigation—that of establishing significant relationships between self-ideal discrepancy and measures of manifest anxiety and observed insecurity—a three-dimensional (2 × 2 × 2) analysis of variance design was used (**10**). The factors controlled in this design were sex, the experimental-control condition, and self-acceptance. The *S*s were divided into two groups according to their scores on the Self-Acceptance Scale—one group (the Highs) made up of those whose Self-Ideal Discrepancy scores fell above the median; the other group (the Lows) with those whose Discrepancy scores fell below the median. This median was computed using all 184 of the main group of *S*s. In order to have equal frequencies in each cell of this design, 20 *S*s were randomly selected from each group which represented a different combination of the factors being controlled; thus, there were 160 *S*s used in the analysis of variance tests. The dependent variables in the analyses of variance were the scores from the Children's Manifest Anxiety Scale and the Kooker Security-Insecurity Rating Scale.

The possible effects of the orientation program on the relationship between self-ideal discrepancy and anxiety was tested by *t* tests of a difference in means between the experimental and control groups.

TABLE 1

TEST-RETEST (ONE WEEK INTERVAL) RELIABILITY CORRELATION COEFFICIENTS OF THE SEVERAL SCORES OF THE SELF-ACCEPTANCE SCALE FOR EACH OF THE TWO CLASSES STUDIED

	N	Self-Concept	Ideal-Self	Self-Ideal Discrepancy
Class I	21	.83	.69	.80
Class II	26	.93	.54	.86

RESULTS

Results of preliminary testing. Test-retest reliability of the Self-Acceptance Scale was determined by administering this test to two classes of sixth graders on two occasions, one week apart. Coefficients of correlation for the two administrations for the various sections of the scale are shown in Table 1.

Although all of these coefficients are significant beyond the one per cent level of confidence, it will be noted that those for the Ideal-Self scores are somewhat lower than the others. This can be attributed to the limited range characteristic of this particular score—that is, the pupils tended not to vary from one another in the rating of their ideal-self concepts—thus, the correlation test was particularly sensitive even to the smallest deviations from one administration to the next.

In order to test whether or not intelligence was likely to be an important factor in the self-acceptance score, IQ scores from the Otis Self-Administering Test of

Mental Ability were secured from school records for all pupils who participated in the testing program. The coefficient of correlation between the Self-Ideal Discrepancy scores and IQ scores for the 47 pupils participating in the preliminary study was −.08 indicating intelligence was not a factor.

Analyses using Self-Acceptance Scale. Tables 2 and 3 show the results of analysis of variance tests using mean scores from the Children's Manifest Anxiety Scale and the Kooker Security-Insecurity Rating Scale, respectively, when the *S*s were divided into two groups, Highs and Lows, according to whether they fell above or below the median Self-Ideal Discrepancy score. These tables show that the differences in the means on both the anxiety and insecurity scales between the Highs and Lows were statistically significant beyond the .01 level of confidence. Furthermore, these differences were in the direction such that the group with the relatively high self-acceptance (that is, low discrepancy scores) yielded average scores indicating less manifest anxiety, as measured, and less insecurity, as rated by observers.

One qualification to these findings should be noted. In the analysis of variance tests, the interaction effects were not significant at the .05 level of confidence (the level prescribed prior to the investigation); however there was a tendency for an interaction to exist between the Self-Ideal Discrepancy scores and the experimental-control condition. (The *F* test of this interaction between the discrepancy scores and the experimental-control condition produced ratios of 3.14 and 3.29 for the Anxiety and Kooker scores, respectively, which are significant between the .10 and .05 levels of confidence.) This tendency towards interaction indicates that the discrepancy measure of self-acceptance was related to the personality variables of anxiety and insecurity differentially according to whether or not the *S*s were in the orientation program. The implications of this possible interaction will be discussed below.

TABLE 2

MEANS AND ANALYSIS OF VARIANCE OF SCORES ON THE CHILDREN'S MANIFEST ANXIETY SCALE OF SUBJECTS DIVIDED INTO "HIGHS" AND "LOWS" ACCORDING TO THEIR SELF-IDEAL DISCREPANCY SCORES ON THE SELF-ACCEPTANCE SCALE

(N = 20 IN EACH SEX, CLASS-TYPE GROUPING)

Self-Ideal Discrepancy Score	Children's Manifest Anxiety Scale			*F*
	Boys	Girls	Total Highs & Lows	
Highs			16.78	13.40**
Exptl. Classes	12.10	18.55		
Control Classes	17.05	19.40		
Lows			12.56	
Exptl. Classes	13.05	13.25		
Control Classes	11.70	12.25		
Total (Boys & Girls)	13.48	15.86		
F	4.30*			

* Significant beyond the .05 level of confidence.

** Significant beyond the .01 level of confidence.

The results reported thus far have indicated relationship between self-acceptance, as measured, and the measures of manifest anxiety and observed insecurity. So as to obtain further information concerning the extent of this relationship, a correlational analysis of the major variables was made, and the results are reported in Table 4. The coefficients of correlation for Self-Ideal Discrepancy scores with the Anxiety scores and Kooker scores were each significant beyond the .01 level of confidence.

Effects of the orientation program. Analysis of the data indicating the effects of the orientation program on the variables meas-

TABLE 3

MEANS AND ANALYSIS OF VARIANCE OF SCORES ON THE KOOKER SECURITY-INSECURITY SCALE OF SUBJECTS DIVIDED INTO "HIGHS" AND "LOWS" ACCORDING TO THEIR SELF-IDEAL DISCREPANCY SCORES ON THE SELF-ACCEPTANCE SCALE

(N = 20 IN EACH SEX, CLASS-TYPE GROUPING)

Self-Ideal Discrepancy Score	Kooker Security-Insecurity Scale			F
	Boys	Girls	Total Highs & Lows	
Highs			25.10	13.42*
Exptl. Classes	24.90	25.65		
Control Classes	25.20	24.64		
Lows			22.81	
Exptl. Classes	22.10	21.60		
Control Classes	24.75	22.80		
Total (Boys & Girls)	24.24	23.68		
F	<1			

* Significant beyond the .01 level of confidence

TABLE 4

COEFFICIENTS OF CORRELATION BETWEEN THE SELF-IDEAL DISCREPANCY SCORE OF THE SELF-ACCEPTANCE SCALE AND THE MEASURES OF ANXIETY AND INSECURITY

	Self-Ideal Discrepancy		
	Boys N = 90	Girls N = 94	Total N = 184
Manifest Anxiety Scale	.28*	.41*	.35*
Kooker Rating Scale	.33*	.34*	.32*

* Significant beyond the .01 level of confidence

ured appears in Table 5. Separation of the experimental classes made possible a comparison of the effects between exposure to the program for one year and for two years. It will be noted that significant differences on the measures of manifest anxiety and observed insecurity appear between the classes having had the program for two years and each of the other two groups involved—those having had the program for one year and those in the control classes. These differences are in a direction which indicates that at least two years' exposure to the orientation program may serve to lower manifest anxiety and reduce observed insecurity. Significant differences between the two-year experimental group and the one-year experimental and control groups were not found with the Self-Ideal Discrepancy scores.

An interesting question is, how do the Highs and the Lows with respect to Self-Ideal Discrepancy in the two-year experimental group compare with those in the control group? This analysis is given in Table 6.

Inspection of this table shows that there is a difference in means between the High two-year experimental group and the control group in manifest anxiety which is significant beyond the .01 level of confidence. In other words, it appears that those in the two-year experimental group who retained a relatively high Self-Ideal Discrepancy appeared to indicate less manifest anxiety (as measured) than did the control *S*s who also had a relatively high Self-Ideal Discrepancy. This lends some support to the contention that at least a two year exposure to the orientation program may allow individuals to feel more comfortable about discrepancies which they feel exist between their self-concepts and ideal-self concepts.

Inspection of Table 6 also reveals that the difference between the Highs and Lows in manifest anxiety in this two-year experimental group is not significant. In other words, it appears that the major results concerning the relationship between Self-Ideal Discrepancy scores and manifest anxiety should be qualified. High Self-Ideal Discrepancy scores may be associated

TABLE 5

MEANS AND *t* TESTS BETWEEN ONE AND TWO YEAR EXPERIMENTAL GROUPS AND BETWEEN THE TWO YEAR EXPERIMENTAL GROUP AND CONTROL GROUP ON MEASURES OF SELF-ACCEPTANCE, ANXIETY, AND INSECURITY

	N	Self-Ideal Discrepancy Score		Manifest Anxiety Scale		Kooker Rating Scale	
		$\overline{X}$	*t*	$\overline{X}$	*t*	$\overline{X}$	*t*
One-Year Exptl. Group	53	7.98	0.87	15.49	2.38*	24.03	2.24*
Two-Year Exptl. Group	45	7.16	0.82	11.93	2.30*	22.42	2.24*
Control Group	86	7.88		14.97		24.12	

* Significant beyond the .05 level of confidence.

TABLE 6

MEANS AND *t* TESTS ON THE CHILDREN'S MANIFEST ANXIETY SCALE BETWEEN THE TWO YEAR EXPERIMENTAL AND CONTROL GROUPS AND BETWEEN THE HIGHS AND LOWS IN SELF-IDEAL DISCREPANCY SCORES

	Children's Manifest Anxiety Scale			
	Highs	Lows	Difference	*t*
Two-Year Exptl. Group	11.88 (N = 16	12.04 (N = 26)	0.16	n.s.
Control Group	18.22 (N = 40)	11.98 (N = 40)	6.24	4.00*
Difference	6.34	0.06		
t	2.92*	n.s.		

* Significant beyond the .01 level of confidence.

with high Manifest Anxiety scores only for *S*s with insufficient self-understanding; this relationship may not apply to *S*s experiencing an orientation program in self-understanding for at least two years. This finding also may explain the tendency for interaction effects to exist between the Self-Ideal Discrepancy scores and the experimental-control condition, particularly in the case of the Anxiety test analysis which was noted above.

DISCUSSION

The validity of the Self-Acceptance Scale. The results reported above indicated significant relationships between the Self-Ideal Discrepancy measure of self-acceptance and scores from the Children's Manifest Anxiety Scale and the Kooker Security-Insecurity Rating Scale. These findings corroborate with sixth-grade-children findings of various other investigators who used similar measures with older *S*s. The validity of the Self-Acceptance Scale is further substantiated by the fact of its relationship to an observation measure (the Kooker scale) which was independent of the children's paper-pencil responses.

The relatively low correlation between the Kooker and the Anxiety scales (N = 184, r = .26) may indicate that these two instruments are measuring different variables, and the Self-Acceptance Scale may be tapping some of the variables distinctive to each of these other two scales as is indicated by the following coefficients of correlation (N = 184): (*a*) Correlation between Self-Ideal Discrepancy scores and Anxiety scores = .35; (*b*) Correlation between Self-Ideal Discrepancy scores and Kooker scores = .32.

The results of this investigation (particularly the analyses of the effects of the orientation program to be discussed in more detail below) raise a serious question as to the adequacy of the self-ideal congruence concept of self-acceptance. A discrepancy between self and ideal might well mean different things to different individuals. While to one person a self-ideal discrepancy might be a threat to his self system, to another such a discrepancy might indicate that his aspirations are high and serve as a challenge to him. What seems to be important is not the discrepancy itself, but the feelings about it. Thus, as the discussion below will indicate, it isn't necessarily the high discrepancy alone which is associated with anxiety but a high discrepancy under certain conditions.

The effects of the orientation program. In the introductory discussion, it was pointed out on an a priori basis that the orientation program might be expected to have diverse effects on the behavior of children. Among the early effects of such a program may be the reduction of the anxiety which seems to accompany a high self-ideal discrepancy. Bearing this out was the tendency for interaction effects to exist, particularly in the analyses of the Anxiety scores, between the Discrepancy scores and the experimental-control condition. This indicates that the relationship between the Discrepancy scores and the measure of anxiety was not the same for the experimental classes as for the control classes.

Further evidence indicating the nature of this tendency for interaction is obtained when analysis is made of the scores of those classes in which most of the pupils had had the orientation program for a two-year period. It will be recalled from Table 6 that those with high Self-Ideal Discrepancy scores in the two-year experimental group had average Anxiety scores which were significantly lower than those in the control group. Thus support is given to the contention that at least a two year exposure to an orientation in self-understanding may allow individuals to feel more comfortable about discrepancies which they feel exist between their self-concepts and ideal-self concepts.

Indication that a period longer than one year in the orientation program may be needed for measurable changes to occur in such variables as manifest anxiety and observed insecurity was seen in the results reported in Table 5. It will be recalled that in this analysis, the two-year experimental group had average scores which indicated less manifest anxiety and less observed insecurity when comparisons were made with the one-year experimental group and the control group. Even though the differences of Self-Ideal Discrepancy scores between these groups were not statistically significant, the discussion above indicated that the feelings about a high discrepancy, where it existed, may have been changed as a result of the two-year orientation program.

Obviously, a single study is not sufficient to establish the validity of this finding. Several studies will be needed to check these results, but those from this investigation—that is, the tendency toward interaction effects noted in the analyses of the Self-Acceptance Scale and the study of the two-year experimental group—suggest that anxieties relative to high self-ideal discrepancies may be more prevalent for those with insufficient self-understanding than for those who have been trained in understanding themselves.

Summary

The purpose of this investigation was to construct a measuring instrument which could be reliably used in the investigation of purported relationships between self-acceptance and other important personality variables and in the study of the possible effects a learning program concerned with the causes of behavior might have on the participants in such a program.

The Self-Acceptance Scale, constructed to measure self-acceptance in sixth grade children, consisted of a series of descrip-

tive statements with which each child rated himself according to the extent he felt he was like the description (the self-concept) and to what extent he felt he wanted to be like the description (ideal-self). Self-acceptance was defined as the congruence (that is, relative lack of discrepancy) between self-concept and ideal-self.

The *Ss* were 184 pupils in eight sixth grade classes in different elementary schools of a medium-sized city in Iowa. Four of the eight classes had undergone a planned learning program designed to help each pupil acquire an understanding of the dynamic, variable, and complex nature of human motivation.

The other measures used in this investigation included the Children's Manifest Anxiety Scale and the Kooker Security-Insecurity Scale (ratings made by trained observers in the classrooms).

The results indicated a statistically significant relationship (beyond the .05 level) between self-acceptance, as measured by the Discrepancy scores, and the measures of manifest anxiety and observed insecurity such that those with the smaller Discrepancy scores had average scores indicating less anxiety and less insecurity. However, a tendency was noted for interaction effects to be present indicating that this relationship between self-acceptance and measures of anxiety and insecurity might have been operating differentially between the experimental and control classes.

When analyses were made of those experimental classes in which most of the pupils had had the program for two consecutive years, support was found for the contention that although Self-Ideal Discrepancy scores remained high for some of these pupils, the Anxiety scores were significantly lower than those of either the one-year experimental or the control classes. This finding suggested that the statement of relationship between self-ideal discrepancy and anxiety might have to be qualified to apply to *Ss* with insufficient self-understanding. Also, the two-year experimental group obtained average scores which indicated less manifest anxiety and less observed insecurity than did the average scores obtained by either the one-year experimental group or control group.

REFERENCES

1. Bills, R. E. A validation of changes in scores on the Index of Adjustment and Values as measures of changes of emotionality. *J. consult. Psychol.*, 1953, **17**, 135–138.
2. Brownfain, J. J. Stability of the self-concept as a dimension of personality. *J. abnorm. soc. Psychol.*, 1952, **47**, 597–606.
3. Bruce, P. A study of the self-concept in sixth grade children. Unpublished doctoral dissertation, State Univer. Iowa, 1957.
4. Castaneda, A., McCandless, B. R., & Palermo, D. The children's form of the Manifest Anxiety Scale. *Child Develpm.*, 1956, **27**, 317–326.
5. Cunningham, Ruth, Elzi, Anna, Farrell, Marie, Hall, J. A., & Roberts, Madeline. *Understanding group behavior of boys and girls.* New York: Bureau of Publ., Teachers College, Columbia Univer., 1951.
6. Jersild, A. T. *Child psychology.* (4th ed.) New York: Prentice-Hall, 1954.
7. Jourard, S. M., & Remy, R. M. Perceived parental attitudes, the self and security. *J. consult. Psychol.*, 1955, **19**, 364–366.
8. Kooker, E. An investigation of security, insecurity, achievement, and boredom in elementary school children. Unpublished doctoral dissertation, State Univer. Iowa, 1951.
9. Levitt, E. E., & Ojemann, R. H. The aims of preventive psychiatry and causality in grade school children. *J. Psychol.*, 1953, **36**, 393–400.
10. Lindquist, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.
11. Ojemann, R. H. An integrated plan for education in human relations and mental health. *J. natl. Ass. Deans of Women*, 1953, **16**, 101–108.

12. OJEMANN, R. H., LEVITT, E. E., LYLE, W. H., & WHITESIDE, MAXINE F. The effects of a causal teacher-training program and certain curricular changes on grade school children. *J. exp. Educ.*, 1955, **24**, 95–114.

13. ROBERTS, G. E. A study of the validity of the Index of Adjustment and Values. *J. consult. Psychol.*, 1952, **16**, 302–304.

14. ROGERS, C. R. The significance of the self-regarding attitudes and perceptions. In M. L. Reymert (Ed.), *Feelings and emotions.* New York: McGraw-Hill, 1950.

15. ROGERS, C. R. & DYMOND, ROSALIND (Ed.) *Psychotherapy and personality change.* Chicago: Univer. Press, 1954.

16. SECORD, P. F. & JOURARD, S. M. The appraisal of body-cathexis: Body-cathexis and the self. *J. cousult. Psychol.*, 1953, **17**, 343–347.

Received April 7, 1958.

Journal of Educational Psychology
Vol. 49, No. 5, 1958

EFFECTS OF TRAINING IN ALTERNATIVE SOLUTIONS ON SUBSEQUENT PROBLEM SOLVING

WALTER I. ACKERMAN
Beth-El Day School, Belle Harbor, Long Island

AND

HARRY LEVIN[1]
Cornell University

An organism's adjustment to a new situation may be scrutinized for the variability or fixedness of the behavior employed. Presumably, during the early stages of solving a novel problem, behavior is characterized by its variability; as commerce with the problem increases and correct solutions are achieved, the resulting habits act to decrease subsequent variability. In a recent statement, Scott (**12**, p. 61) points out that such a sequence is characteristic even of the simplest organism:

> Such variability can be seen even in the lower organisms. If a paramecium runs into an obstacle it does not keep repeating its behavior. It backs off and approaches from a different angle, and never does exactly the same thing again. It is apparent that variability of this type is necessary for the process of adjustment, since an animal which gave fixed invariable responses could never adapt itself to a variety of changing conditions.

The problem solving model presented by the paramecium might well be envied by humans. There is ample experimental evidence that under some circumstances human problem solving is characterized by fixedness rather than variability of response. If the demands of the problem situation are similar to previously acquired solution methods, the transfer effects to novel situations should, of course, be positive. Where the demands of the problem are only superficially similar to previous successful solutions the present application of these methods is often doomed to failure. The next adaptive step would be to abandon the first method, or hypothesis, and to try another, in contrast to the persistent application of the first.

Fixedness in problem solving has been variously attributed to immediately prior problem solving experiences, to the nature of the problem at hand, and to personality dispositions of the solver (**1, 4, 5, 8**).

One major cause of lack of variability is problem solving set. Of course, should the person's set be appropriate for the problem the solution is likely to be facilitated. Under the not infrequent conditions where his set is inappropriate, the problem solver has a dual task before him, ridding himself of the maladaptive set and then applying the new solution. Recognizing Gibson's (**3**) indictment against the chaotic usage of the concept of set, we will define it for our purposes as "that manner of attacking a problem which is carried over from a previous to a succeeding problem situation."

Set precludes the variability of behavior which is essential if the habituated tack is to be cast off. This study is concerned with a way of training children

[1] This study is part of an Ed.D. thesis submitted to the Graduate School of Education, Harvard University, by the first author. The research was performed when both authors were members of the Harvard Teacher Education Research Project, supported by grants from the Fund for the Advancement of Education.

in problem solving so that in the face of new problem situations they will either not develop sets or will give them up when they do not work.

What the problem solver needs is a nonspecific approach, (sometimes called also "mode of attack" or "plan of action"), to many types of problems. One important aspect of this general approach is the assumption that problems often lend themselves to various solutions, so that if one solution does not work there is still the possibility that the problem may be solved. The solver might profitably attack the problem again, with a new orientation. How may such general problem solving behavior be inculcated? Maier (**7**) gave *S*s specific training in reasoning and Luchins (**6**) prior to his einstellung situation instructed his *S*s, "Don't be blind." Both investigators report that these devices aided *S*s in overcoming habituated ways of solving problems.

This study investigates another method of inducing variability in problem solving. One group of *S*s is taught two solutions to the same set of problems, another group one solution to these problems. Then the two groups are observed on a set of similar problems necessitating new solution methods. Next, a set is induced on a novel series of problems and the responses to situations where this set no longer works are observed. Our expectations are that the people trained in alternative solutions will solve more test problems correctly, will exhibit greater variability to these problems, and will persevere longer on problems too difficult for them to solve.

Method

The *S*s were 48 sixth grade children, 25 boys and 23 girls, who lived in a predominantly middle class, white collar community.

*S*s were divided into two groups of 24 children each, with the sexes represented as equally as possible in each group. The two groups did not differ either in intelligence or school achievement. *E* took each child individually from his classroom to a nearby experimental room. *E* told the child that he was interested in "how children solved problems." The procedure consisted mainly of training and test series on two separate types of paper and pencil problems.

Problem 1. The water jar problems. These problems are an adaptation of the water jar problems described by Luchins (**5**). *S* is required to obtain a specified amount of water using only empty jars and their total capacities as measures. The jars were drawn on cards, with a number in each jar denoting its capacity and a number in the margin indicating the amount of water to be obtained. *E* presented a single problem on a card, and *S* solved the problem on a work sheet.

The training series consisted of 10 problems, each solvable by the formulas, B − A − C, or B − 2A + C (the letters refer to the capacities of the three jars). *E* taught *S* to solve the problems by one or both of these solution methods depending on the experimental design below.

Following the training problems and a repetition of the general instructions, *S* was presented six test problems. Each of the first five called for a novel solution method; the sixth was solvable by the method taught during the training series.

S was permitted to work on each test problem until he solved it or until he decided to stop. The time spent on each problem was recorded.

Problem 2: The puzzle problems. Immediately following the completion of the above training and test series the *S*s were given 13 puzzle problems. These are paper and pencil versions of jigsaw puzzles. *S* was given a booklet on each page of which was a picture of a whole figure. Beneath the figure were pictured numbered fragments of various sizes and shapes. The problems call for *S* to choose those fragments which, when put together, will form the whole figure. *S*s indicated their solutions at the bottom of the page by marking the numbers of the pieces they chose, in the order in which they used them.

The first eight of the 13 puzzles were all solvable by assembling the fragments numbered 1, 3, 5, and 7. Each of the five final problems required a different solution from the initial set-inducing series and from each other. Superficially, the test problems looked as though they were solvable by the original method. To the *S*, there was no break between the training and the test problems.

No time limit was set on these problems; a record of the time spent on each problem was kept.

Experimental manipulations. The two groups of *S*s differed only in the treatment they received during the training series of 10 water jar problems. One group was taught the two alternative solutions to these problems (labelled A group). *E* worked the first problem by one and then the other solution method and half of the succeeding problems were worked out by one or the other method. The no alternatives group (hereafter labelled NA) were taught to solve the 10 problems by only one solution method.

Results

Stated generally, our hypotheses are (*a*) that experience in solving problems in more than one way should dispose the *S* to relinquish more quickly a maladaptive set and consequently yield more correct solutions to new problems, and (*b*) that this variability of response should generalize to problems different from those on which the dual solutions were taught.

Consider first the six test problems in the water jar series. Group A solved 111 problems correctly, whereas Group NA produced 94 correct solutions. Though the direction of difference is as perdicted, the difference is not statistically significant. More critical than the total number of correct solutions is the behavior on the first test problem. A correct answer to this problem indicates to the *S* that the solution methods of the training period may not be applicable to the ensuing problems and hence may aid him in overcoming whatever set was induced by the earlier series. Group A achieved 12 correct answers; group NA, 9. Here too the direction is interesting, but the differences are not significant.

The results are clearer when we compare the two groups in the number of other-than-training solutions they offered to the six test problems. Since there is abundant evidence that the sexes differ in their problem solving behavior, "sex" was added to "treatment" as a criterion of classification in all the analyses of variance reported.[2] As can be seen in Tables 1 and 2, Group A offered significantly more "other than training" solutions than did Group NA. We may infer then that training in alternative solutions leads to more variable problem solving behavior, though it is not clear that the increased variability yields more correct solutions.

No time limit was set on any water jar problem. The instructions directed the *S*s to work on a problem until either a satisfactory answer was achieved or until he thought that it was no longer profitable to continue on that problem. The amount of time spent on a problem before giving up was, then, a measure of perseverance. Table 3 gives the number

[2] A correction factor for the unequal cell frequencies was applied to all of the analyses of variance.

TABLE 1
MEAN NUMBER OF OTHER-THAN-TRAINING SOLUTIONS TO TEST WATER JAR PROBLEMS

	Group A	Group NA
Male	9.33 (12)	7.15 (13)
Female	11.42 (12)	7.91 (11)

Note.—Numbers in parentheses are the numbers of *S*s in each cell.

TABLE 2
SUMMARY OF ANALYSIS OF VARIANCE OF OTHER-THAN-TRAINING SOLUTIONS TO TEST WATER JAR PROBLEMS

Source	*df*	Sum of Squares	Mean Square	*F*
Sex	1	2.03	2.03	1.22
Treatment	1	8.09	8.09	4.87*
Sex × Treatment	1	0.44	0.44	
Error	44	73.09	1.66	

* $P < .05$.

TABLE 3
MEAN PERSISTENCE (IN SECONDS) ON TEST WATER JAR PROBLEMS

	Group A	Group NA
Male	343.43 (7)	178.11 (9)
Female	131.41 (10)	101.84 (8)

of *S*s in each sex and group who gave up the problems before actually achieving a correct solution and the time in seconds spent on the problems before giving up. The analysis of variance summarized in Table 4 indicates that boys persevered longer than girls and that Group A persevered longer than Group NA. The greater perseverance of boys confirms other findings about sex differences in perseverance (**10, 11**). The training in multiple possibilities of solving a problem likely taught the NA group that failure with one method did not exhaust the possibilities of success, so that they may profitably spend additional time on the problem. This finding complements Robinsons' (**9**), who reports that *S*s who thought their chances of arriving at a correct solution to a problem were excellent spent more time on that problem than did *S*s who were less confident.

The first part of the experiment confirms several of our expectations. There was a tendency for the group trained in alternatives to solve more problems correctly. This group offered a greater variety of solutions to the test problems, and on those problems they could not solve they worked longer before they gave up.

The puzzle problems were introduced to test whether the effects of training in alternative solutions would transfer to a

TABLE 4
SUMMARY OF ANALYSIS OF VARIANCE OF PERSISTENCE ON TEST WATER JAR PROBLEMS

Source	*df*	Sums of Squares	Mean Square	*F*
Sex	1	9495.53	9495.53	7.55*
Treatment	1	20777.78	20777.78	16.53**
Sex × Treatment	1	4607.02	4607.02	3.66
Error	30	37,705.80	1256.86	—

* $P < .01$.
** $P < .001$.

novel set of problems. Each of the first eight puzzles were solvable by the use of, in each case, pieces numbered 1, 3, 5, 7. The succeeding five puzzles required different pieces for their correct construction. Does training in an earlier, different type of problem influence behavior on the new test problems?

As with the water jar problems, the A group reached more correct solutions than did their controls. They solved 61 problems correctly; Group NA solved 49. On the first test problem, there were four correct solutions in the A group and none in the NA group. Though neither of these differences is statistically significant, their direction was predicted.

If the set induced by the initial eight jigsaw problems is effective, the *S*s should continue to use the same numbered pieces in the test problems, where they are now inappropriate. Table 5 gives the mean numbers of pieces with the same numbers as those used in the training puzzles for each group. The analysis of variance for these data is summarized in Table 6. The A group evidenced greater variability by using fewer incorrect pieces than did the NA group. This finding permits us to conclude that the ability to overcome inappropriate problem solving sets is acquired through alternative training in solution methods to the same problems and that this ability is transferable to problem situations other than the one on which the alternative training was acquired.

TABLE 5

MEAN NUMBER OF SAME-AS-TRAINING PIECES (PERSISTENCE) USED IN SOLVING PUZZLES

	Group A	Group NA
Male	2.33 (12)	3.08 (13)
Female	1.83 (12)	3.18 (11)

TABLE 6

SUMMARY OF ANALYSIS OF VARIANCE OF NUMBERS OF SAME-AS-TRAINING PIECES USED IN SOLVING PUZZLES

Source	*df*	Sum of Squares	Mean of Squares	*F*
Sex	1	.04	.04	—
Treatment	1	1.10	1.10	12.37*
Sex × Treatment	1	.09	.09	—
Error	44	3.91	.089	—

* $P < .001$.

Discussion

Some impressions about the water jar problems peripheral to our interests, but germane to Luchins' (**5**) results, occurred to us during the experiment. Luchins, it will be recalled, creates a set to use combinations of three water jars during his training problems and during the test series many *S*s are not able to use simpler, two jar solutions. The interpretation is that set precludes *S*s from seeing the easier solution method. In some instances in the present study it was not that the *S*s did not see the simpler solution, but rather they did not understand that they were permitted to use two jars. Some of our *S*s actually tried and abandoned correct two jar solutions because, as they explained it, they thought that since the initial problems involved three jars, the continued use of all three was a part of the task. It may be therefore, that in the previous water jar studies, there were people who were actually aware of the simpler solution but were not clear that it was acceptable.

The pedagogical implications of our findings are apparent. Although it involves an extrapolation from laboratory to classroom conditions, we might expect that a teacher's consistent training in alternative solutions to problems might result in efficient overcoming of set.

Put into historical perspective, the findings of this study are congruent with Woodworth's general factors theoiy of transfer. Apparently, what occurs when the child is given even brief training in solving problems by more than one method is that he develops a "general approach to novel problems." Colloquially, this approach is self-instruction to abandon an inefficient solution method and to try something new. One wonders whether this general problem solving skill was verbalized by our experimental *S*s or whether it operated as an unverbalized, functional concept. Although we have no data on this point, we might expect that the verbalization of the principle would increase its efficiency.

Summary

Two groups of sixth grade children were trained to solve 10 water jar problems. One group was given a single solution to the problems; the second was taught that the problems were solvable by either of two methods. On a succeeding set of test problems which necessitated solutions different from the two training methods, the alternative group tended, though not significantly, to solve more problems correctly, offered significantly more other-than-training solutions, and persevered longer on problems they were unable to solve.

To test the transfer effects of the original training, *S*s were given 13 jigsaw puzzles: the first eight were all solvable by a single method so that a set might occur; the final five required various solution methods. Here, also, the *S*s who were trained in alternative methods on the earlier water jar problems evidenced greater problem solving variability by using significantly more pieces of the puzzles which were not used in the training series.

References

1. Adamson, R. E. Functional fixedness as related to problem solving. *J. exp. Psychol.*, 1952, **44**, 288–291.
2. Birch, J. G., & Rabinowitz, H. S. The negative effect of previous experience on productive thinking. *J. exp. Psychol.*, 1951, **41**, 121–125.
3. Gibson, J. J. A critical review of the concept of set in contemporary experimental psychology. *Psychol. Bull.*, 1941, **38**, 781–817.
4. Guetzkow, H. An analysis of the operation of set in problem-solving behavior. *J. Gen. Psychol.*, 1951, **45**, 219–244.
5. Luchins, A. S. Mechanization in problem solving. *Psychol. Monogr.*, 1942, **54**, No. 6 (Whole No. 242).
6. Luchins, A. S. Classroom experiments on mental set. *Amer. J. Psychol.*, 1946, **54**, 295–298.
7. Maier, N. R. F. An aspect of human reasoning. *British J. Psychol.*, 1933, **24**, 144–155.
8. Maier, N. R. F. The behavior mechanisms concerned with problem-solving. *Psychol. Rev.*, 1940, **47**, 43–58.
9. Robinson, Elsa E. An experimental investigation of two factors which produce stereotyped behavior in problem situations. *J. exp. Psychol.*, 1940, **27**, 394–410.
10. Ryans, D. G. The measurement of persistence. *Psychol. Bull.*, 1939, **26**, 715–739.
11. Schofield, W. Jr. An attempt to measure persistence in its relationship to academic achievement. *J. exp. Psychol.*, 1943, **33**, 440–445.
12. Scott, J. P. The genetic and environmental differentiation of behavior. In D. B. Harris (Ed.), *The concept of development.* Minneapolis: Univer. Minnesota Press, 1957.

Received April 17, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 5, 1958

THE EFFECTS OF DIFFERENT TEACHING METHODS: A METHODOLOGICAL STUDY[1]

MARVIN NACHMAN[2] AND SEYMOUR OPOCHINSKY

University of Colorado

Reviews of teaching research have consistently concluded that different teaching procedures produce little or no difference in the amount of knowledge gained by the students (**1, 2, 3, 5, 6**). This same conclusion has been reached despite the fact that experimenters have employed a wide variety of independent variables, such as lecture versus discussion classes, instructor-centered versus student-centered classes, large versus small classes, various types of TV classes, etc. These results are surprising if one considers that much of the research was instigated by the hypothesis that differences would be found. Furthermore, it appears as if most educators still assume that classroom techniques do in fact have specific effects. Why then have differences not been found?

One obvious hypothesis is that the teaching methods which have been employed are not sufficiently distinct to produce significant differences in the amount of knowledge acquired. The purpose of this paper is to examine an alternative hypothesis, namely, that the different teaching methods have, in fact, produced differential amounts of learning but that these effects have been masked in the measurement process.

Typically, in measuring the effectiveness of different teaching methods, one of the major dependent variables has been the performance of the students on the final examination. It is clear that variance on the final examination is due to many factors in addition to the specific teaching methods employed. Such things as the intellectual ability of the student, his motivations, the amount of studying he has done outside of class, various personality factors and environmental pressures, etc. will also affect his performance. In research on teaching methods, most of these other variables are not controlled (usually, the major control is to equate students for ability) and, since they undoubtedly account for a significant proportion of the variance, it is perhaps not surprising that the diverse experiments on teaching methods have so consistently concluded that there are no significant differences as a function of teaching method employed.

A basic difficulty of this conclusion stems from the fact that students prepare for examinations by studying outside of class. It is perfectly possible that two groups taught under different procedures may learn very different amounts in class, but when both then engage in significant extra amounts of study, the difference in performance becomes negligible. This is especially true since most final examinations are based, at least in part, on information to be found in the student's textbook. Thus, if both groups study the textbook equally, one might expect that they will do about equally well on the examination even though one group has learned more in class. The matter can, of course, be more complex. It may be, for example, that if Group A learns more in class than Group B, then Group B feeling itself less well prepared for the pending examination will study more than will Group A. That is, one might find that students in preparing for an examination, study until they feel they know the material to a certain degree, and if one group learns the material in class they will feel less need for addi-

[1] A portion of this paper was presented at the 1958 meeting of the APA.

[2] Now at the University of California, Riverside.

tional study than the group which has not learned the material in class.

This discussion suggests that to evaluate a particular teaching method effectively, testing of the students ought to occur immediately after the method has been employed thereby avoiding contamination by the additional variable of outside study. This procedure would be a more direct test of the influences of the teaching situation *per se* and would therefore be more likely to reveal its effects, even if these effects are not observable on a final examination.

The following experiment was designed to test this general prediction. The independent variable of class size was employed and it was hypothesized (*a*) that students in a small class would do better than students in a large class on quizzes for which they had not prepared, and (*b*) that no differences would be found between the classes when students were given an opportunity to study for an examination. The experiment was not specifically concerned with the variable of class size. Rather, it was used because it offered a convenient and simple way of testing the hypotheses and because it has so consistently been shown to have no influence on the amount learned by students when traditional measurements have been employed. In one review on the effects of class size, the author concludes that there have been "... more than 200 studies which clearly reveal that there are no consistent differences" (**4**).

Method

Subjects

Two sections of about 150 students each were enrolled in the senior author's General Psychology course. The day after the students returned from their Christmas vacation, they were informed that for the last two weeks of the semester the instructor intended to meet a small class which would be identical in all respects to the larger classes except for size. The class was arranged for Monday, Wednesday, and Friday at 8:00 A.M. and 42 of the 300 students volunteered to attend that class rather than their large one. These students were divided into two groups of 21 students and each student in one group was very closely equated with a student in the other group on the basis of the three hourly examinations they had previously taken in the course that semester. By a flip of a coin, one group of 21 was then selected for the small class and the other group remained in one of the two large classes. Thus, the *S*s for the experiment consisted of two equated groups of 21 students each, one group in a small class of 21 students and the other group in large classes of about 140 students.

Procedure

Since the experiment was not concerned with testing the unique advantages of a small class versus a large class, the teaching method used in the different classes was as nearly identical as possible. The small class met on MWF at 8:00 A.M. and the two large classes met on MWF at 10:00 and 11:00 A.M., respectively. It was arranged that the small class meet earlier in the day than the large classes so that, in the event the students communicated information to each other about examinations, the bias introduced would operate against the hypotheses. (Questioning of the students after examinations revealed that the amount of communication was just about nil.)

All three sections were conducted by the lecture method and, as much as possible, the lecturer repeated exactly the same material to the three classes. During the last 10 minutes of the second and fourth lectures, the students in all three classes were given a "pop-quiz" which specifically covered the material that was presented in that lecture and the previous lecture. The students had typically received a few weeks notice for examinations before this time and were clearly not expecting and had not prepared for these quizzes.

In order to avoid biasing the lectures in favor of the quiz material, the quizzes were constructed by the junior author and were not seen by the lecturer until after they were administered. Both quizzes were multiple choice, the first containing 15 questions and the second 8 questions. (The second quiz was originally designed as 15 questions but since only 8 of the items were actually covered in the lectures, the rest were discarded before the quizzes were scored.)

The students took the final examination about one week after the last day of classes. The final examination consisted of 125 multiple choice items and was made up by several members of the department. It was administered at the same time to about 1,000 students, of whom 42 were the *S*s for this experiment. Twenty-five of the items on the final examination covered the same material which had been presented while the students were in the large versus small class experimental situation. The two groups of 21 students were compared in their performance on the two quizzes and on these 25 items of the final examination.

Results

The two quizzes and the final examination were scored by assigning one point for each correct answer. For each student, a total quiz score was obtained by adding the scores of his two quizzes, and difference scores were then computed for each matched pair of students in the small and large classes. Because 13 of the 42 students were not present for one or the other of the two quizzes (seven students from the large classes and six students from the small class), the difference scores for these students were based on the one comparable quiz which both members of the matched pair had taken. The difference scores revealed that 13 of the small class students did better than their matched pairs in the large classes whereas only 3 did worse and there was no difference for 5 students. The differences analyzed with a *t* test for matched pairs resulted in a *t* of 2.57, which for 20 degrees of freedom has a probability level of less than .02.

An alternative method of analyzing the results was to correlate the two quizzes (the Pearson product-moment coefficient was .56) and to use the regression equation to predict the scores on the quizzes missed by the absent students. This resulted in two quiz scores for each student which were added. The mean of this sum was 16.02 for the small class as compared to 14.34 for the matched students in the large classes. The standard error of the difference between the means was .56 which resulted in a *t* of 3.00 which for seven degrees of freedom (since 13 of the scores were predicted) also has a probability level of less than .02.

On the 25 items of the final examination, 10 of the small class students did better than their matched pairs in the large classes while nine did worse, and there was no difference for two students. The mean for the small class was 18.00 as compared to 17.53 for the matched students in the large classes. The standard error of the difference between the means was .90 which resulted in a *t* of .53 which is not significant.

Discussion

The results confirmed the hypotheses that differential performance would be found on quizzes which specifically covered classroom material and for which the students had not prepared but that performance would be equal on final examinations for which the students had devoted a large amount of extra study. One cannot be certain, of course, about the role attributed to extra study. The specific conclusion which can be drawn from the data is that in order to test the effect of a particular classroom technique, evaluation should be done immediately after the technique is employed. Waiting until a final examination (or any other announced examination) confounds the problem by permitting many other variables to operate, one of the most obvious of which is extra study.

In most experiments on learning, when one is measuring the effects of a particular variable, e.g., massed versus distributed practice, the learning opportunities of the *Ss* are limited almost exclusively to the experimental situation. This has not been the case in experiments on teaching methods, perhaps because of the difficulties of controlling some of the extraneous variables. It may be, however, that many of the experiments on teaching methods would have led to significant differences if the evaluation procedure avoided contamination by such factors as extra study. The fact that in previous studies, the variable of class size has so repeatedly been found to produce no significant effects on amount learned, and yet yielded significant differences in the present experiment, implies that a restudy of other teaching variables using different measuring techniques might be more fruitful than it has previously been.

Although the experiment was not primarily concerned with the effects of small versus large classes, the data indicate that students in the small class learned significantly more in class than did the students in a large class. The difference of about 1.7 points on a test which had a mean score of about 15 was fairly large considering the relatively small variance of the difference scores and the fact that the small versus large class variable was limited to only two lectures per quiz. Furthermore, there was no attempt to utilize the unique advantages of a small class by permitting more questions or discussions.

What was the specific variable operating in the large and small classes which produced the difference? As Buxton (**2**) has pointed out, class size, like time, is not a variable by itself but rather is an abstraction in which other variables may be permitted to operate. In evaluating the course at the end of the semester, almost all of the students in the small class spontaneously commented that they found it easier to pay attention in the small class and that they had become more interested. Casual observations by the lecturer corroborated this. The students in the small class appeared to be much more alert in their listening as well as note-taking behavior. Rarely, if ever, were they observed in behaviors which are not uncommon in very large classes, such as talking to each other, staring out the window, reading, etc. Undoubtedly, the very proximity of the lecturer in the small class acts as an inhibitory influence on these behaviors.

It is also possible, of course, that the variable of class size had nothing to do with the obtained differences. It may be that something like a "Hawthorne-effect" was operating in which the students in the small class felt more significant and more highly motivated because a special class had been established for them. This possibility, and similar ones, such as unintentional lecturer bias, or that students perhaps learn more at 8:00 A.M. than at 10:00 or 11:00 A.M., do not reduce the significance of the major finding, namely that differences as a result of teaching technique manipulations were found on "pop quizzes" but not on a prepared-for final examination.

Summary and Conclusions

Twenty-one students in a small class were compared on examination performance with a matched group of students who were in a large class. It was hypothesized that the small class would do better on quizzes which specifically covered the

classroom material and for which the students had not prepared but that the two groups would do equally well on final examinations for which they had studied. The hypothesis was confirmed and the implications of this methodological procedure were discussed in relation to other research on the effectiveness of different teaching methods.

REFERENCES

1. Birney, R., & McKeachie, W. The teaching of psychology: A survey of research since 1942. *Psychol. Bull.*, 1955, **52**, 51–68.
2. Buxton, C. E. *College teaching, a psychologist's view.* New York: Harcourt-Brace. 1956.
3. Good, C. V. Colleges and universities—VIII. Methods of teaching. *Encyc. educ. Res.*, 1950, **42**, 273–278.
4. Kidd, J. W. The question of class size. *J. higher Educ.*, 1952, **23**, 440–444.
5. Otto, H. J., & von Borgersrode, F. Class size. *Encyc. educ. Res.*, 1950, **42**, 212–215.
6. Wolfle, D. L. The first course in psychology. *Psychol. Bull.*, 1942, **39**, 685–712.

Received April 17, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 5, 1958

CURRICULAR DIFFERENCES IN JOB INCENTIVE DIMENSIONS AMONG COLLEGE STUDENTS

A. W. BENDIG AND EUGENIA L. STILLMAN
University of Pittsburgh

In a recent study (**1**) the results of a preliminary attempt to identify the dimensions of job incentives among college students was reported. By means of a factor analysis of rankings of a homogeneous list of job incentives by college *S*s three dimensions were isolated and tentatively identified as: (*a*) need achievement vs. fear of failure; (*b*) interest in the job itself vs. the job as an opportunity for acquiring status; and (*c*) job autonomy of supervision vs. supervisor dependency. Simple procedures were developed for computing factor "scores" for individual *S*s from the differences between pairs of ranked incentives.

If the dimensions underlying the ranked incentives are related to choice of occupational goal among college *S*s, we would expect that *S*s divided into more homogeneous subgroups on the basis of their curricular courses of study would show somewhat greater agreement in their ranking of the incentives than the combined curriculum heterogeneous total group of *S*s and that the factor scores would be capable of demonstrating significant differences among the curriculum subgroups. Curriculum subgroup differences would have the added advantage of helping to define the meaning of the job incentive dimensions.

The present research was designed to (*a*) identify, on an a priori basis, curricular subgroups of *S*s, (*b*) combine these subgroups into major curriculum groupings by means of an empirical factor analysis of their subgroup profiles in ranking job incentives, and (*c*) test whether there were significant differences among these major curriculum groupings on the derived factor scores.

PROCEDURE

A form was prepared which listed the eight incentives used previously (**1**) and it was administered to 267 undergraduate college *S*s (174 men and 93 women) enrolled in 10 sections of undergraduate psychology courses. The form requested the *S* to indicate his (her) name, age, sex, school, and curriculum group within the university, major subject and to write a brief description of the job he expected to accept after graduation. The *S* was then requested to *rank* (from one to eight) the following list of incentives with the incentive that would be most important to him in selecting the job previously described being ranked "one" and the least important incentive being ranked "eight." The incentives used were:

1. Opportunity to learn new skills
2. Friendly fellow workers
3. Freedom to assume responsibility
4. Good job security
5. Good prospects for advancement
6. Full insurance and retirement benefits
7. Recognition from supervisors for initiative
8. Good salary

The 267 *S*s who completed this form were dichotomized as to sex and further divided into eight major curriculum areas:

Business Administration: 51 men.
Engineering and Mines: 42 men.
College B.A. (Humanities and Social Sciences): 35 men and 11 women.
College B.S. (Natural Sciences): 30 men and 11 women.
Pre-Education: Elementary; 26 women.
Pre-Education: Secondary; 16 men and 25 women.
Pre-Nursing: 11 women.

Nursing Education: 9 women.

This division by sex and curriculum resulted in 11 curriculum subgroups.

Two sets of "scores" were available for each *S*. The eight rankings of the incentives constituted an incentive profile for the *S* and these rankings were averaged for each of the 11 subgroups. A second set of factor "scores" had been developed in the previous study (**1**). The Factor A score was defined as the ranking of Incentive 6 minus the ranking of Incentive 1. Similarly, the Factor B score was Incentive 8 minus Incentive 2 and Factor C was Incentive 7 minus Incentive 4. Positive scores computed in this manner indicate (*a*) high need achievement, (*b*) high interest in the job itself, and (*c*) high need for autonomy from supervision. Negative factor scores presumably measure (*a*) high fear of failure, (*b*) strong attitude toward the job as a stepping-stone for advancement, and (*c*) high need for a dependency relation to the job supervisor.

Results

The average (mean) incentive profiles for each of the 11 curriculum subgroups can be found in Table 1. From the sums of ranks of the eight incentives for each subgroup the average rank-difference (rho) intercorrelation among the *S*s constituting each subgroup was computed by the usual formula (**2,** p. 421) and these coefficients are given in the last column of Table 1. Nine of the 11 coefficients are significantly different from zero at the .01 level of confidence with the remaining two coefficients being significant at the .05 level. The average rho intercorrelation for the total group of 267 *S*s (ignoring curriculum subgrouping) was .20 (significant at the .01 level). The median average intercorrelation of the 11 subgroups in Table 1 is .28 which can be compared to the average intercorrelation of .20 when the curriculum subgroups are combined, indicating that the division of *S*s into curriculum subgroups did somewhat increase the homogeneity with which the *S*s ranked

TABLE 1

Mean Ranks of Incentives and Average Intragroup Correlations (Rho) for Curriculum Subgroups

Curriculum Subgroups	Sex Group	Number of Students	Incentives								Average Intercorrelation
			1	2	3	4	5	6	7	8	
Business Administration	M	51	5.6	5.3	4.6	4.1	1.8	6.7	4.9	3.0	.38**
Engineering & Mines	M	42	4.2	4.8	4.8	4.3	2.6	7.0	5.1	3.2	.26**
College B.A.	M	35	5.3	5.1	4.1	3.8	3.2	6.3	5.3	3.0	.21**
College B.A.	F	11	4.1	3.9	2.7	4.4	4.6	6.8	4.5	5.0	.14*
College B.S.	M	30	4.6	5.4	3.1	2.8	3.5	6.5	5.9	4.2	.28**
College B.S.	F	11	2.6	5.0	3.5	3.5	3.9	7.5	5.5	4.4	.32**
Pre-Education: Elementary	F	26	3.7	4.0	3.1	2.6	5.4	6.0	6.3	4.9	.28**
Pre-Education: Secondary	M	16	4.8	4.5	4.2	3.1	3.7	6.4	4.9	4.4	.10*
Pre-Education: Secondary	F	25	3.7	4.3	2.7	2.6	5.3	7.4	5.4	4.6	.39**
Pre-Nursing	F	11	3.9	3.5	2.5	4.3	3.7	6.8	6.6	4.7	.33**
Nursing Education	F	9	2.3	6.2	2.6	4.4	3.3	7.2	5.7	4.2	.46**

* $P = .05$.
** $P = .01$.

the incentives. The median intercorrelation for the five male curriculum subgroups was .26, while the median coefficient for the six female subgroups was .32, suggesting that the female *S*s were slightly more homogeneous in ranking the incentives than were the male *S*s. However, the 11 coefficients were ranked and the Mann-Whitney U test (**4**, pp. 116–127) was used. The two-tailed probability was approximately .21 which indicates that the hypothesis of similar homogeneity among men and women *S*s in ranking the incentives cannot be rejected. Two of the three smallest average intercorrelations were contributed by the college B.A. subgroups and it is suggested that in future curriculum research the two major samples of *S*s within these subgroups, majors in the humanities and in the social sciences, be treated separately in finding incentive profiles. The intercorrelation among the Pre-Education–Secondary men might be increased by splitting these *S*s in future samples into those majoring in physical education and *S*s majoring in more academic high school subjects.

In order to combine the 11 curriculum subgroups into more meaningful major groupings by means of factor analysis the eight incentive mean ranks for each of the 11 curriculum subgroups were ranked and the intercorrelations among the 11 sets of ranks were computed by the rank-difference method. The mean correlation among the subgroups was .60 with individual coefficients ranging from .01 to .98. Nineteen of the 55 intercorrelations were significant at the .05 level of confidence. This matrix was factor analyzed by the usual centroid method and three orthogonal factors were extracted. The median absolute value of the residuals after extraction of the third factor was only .03, indicating the absence of a fourth factor. The original factors were rotated to simple structure and the rotated orthogonal factor loadings for the curriculum subgroups can be found in Table 2.

The 11 curriculum subgroups appeared to condense into six major groupings on the basis of their patterns of loadings on orthogonal Factors X, Y, and Z, and their factor pattern groups (I to VI) are indicated in Table 2. Factor X, which accounts for 40 per cent of the variance among the subgroups, is obviously a sex factor with the male subgroups all having high loadings on this factor and the female subgroups showing moderate or low load-

TABLE 2

FACTOR ANALYSIS OF RANK INTERCORRELATIONS AMONG CURRICULUM SUBGROUPS

Factor Pattern	Curriculum Subgroups	Sex Group	Number of Students	Factor Loadings			h^2
				X	Y	Z	
I	College B.A.	M	35	96	12	13	95
	Business Administration	M	51	94	−02	10	89
II	Engineering & Mines	M	42	74	09	58	89
III	Pre-Educ: Secondary	M	16	90	42	00	99
	College B.S.	M	30	80	54	14	95
IV	Nursing Education	F	9	37	44	71	83
	College B.S.	F	11	37	66	58	91
V	Pre-Educ: Secondary	F	25	35	89	13	93
	Pre-Educ: Elementary	F	26	33	86	13	87
	Pre-Nursing	F	11	31	76	11	69
VI	College B.A.	F	11	00	91	00	83

ings. Factor pattern Group VI is similar to Group V except for an extremely low loading on the "masculinity" Factor X which appears to justify the separation of Group VI (College B.A. women) from the three female subgroups comprising Group V. Comparing the men and women *S*s on the mean ranks they assigned to the individual incentives indicated that the men preferred Incentives 5 (Good prospects for advancement) and 8 (Good salary), while the women reported that Incentives 1 (Opportunity to learn new skills) and 3 (Freedom to assume responsibility) would be more important in their job selection.

Factor Y (37 per cent of the intersubgroup variance) appeared to be almost the reverse of the "masculinity" Factor X with the exception that the male subgroups within Group III (Pre-Education: Secondary and College B.S.) had moderate loadings on Factor Y (median loading = .48) that were intermediate in size between the loadings of the six female subgroups (median loading = .81) and the remaining three male subgroups (median loading = .09). The factor pattern groups were trichotomized as to their loadings on Factor Y with Groups V and VI being high (median = .88), Groups III and IV being intermediate (median = .49) and Groups I and II having low loadings (median = .09). Incentives 3, 5, and 8 appeared to be related to Factor Y with subgroup preferences for Incentive 3 being positively correlated with Factor Y loadings and Incentive 5 and 8 preferences being negatively related to the loadings. Factor Y might be labelled a "tenderminded social service" vs. a "toughminded practical" dimension from an inspection of the curriculum subgroup loadings along this factor. However, "naming" of this factor is less important than is the use of the factor loadings in empirically clustering the subgroups.

Factor Z (12 per cent of intersubgroup variance) was a triplet factor with the three subgroups included in Factor Pattern Groups II and IV showing large loadings and the remaining eight curriculum subgroups showing zero loadings. Groups II and IV seem to prefer Incentive 1 more than did the remaining groups, but comparisons of the high and low Factor Z groups showed slight differences on their mean rankings of the other incentives. We might suggest that Factor Z represents an "interest in science and technology" factor, but again it should be emphasized that identification of this factor by a name was not an aim of this study.

Three factor scores had been computed for each of the 267 *S*s in the 11 curriculum subgroups and three analyses of variance, one for each factor score, were computed to test for significant differences among the subgroups. The total variance ($df = 266$) of a factor score was first split into the variance between the means of the 11 subgroups ($df = 10$) and the residual individual differences variability among the *S*s within the subgroups ($df = 256$). The 11 within subgroup variances were tested for homogeneity by Bartlett's chi-square technique before pooling and the chi-square values gave no evidence of significant heterogeneity for any of the three factor scores. The between subgroups variance was further divided into two components: between factor pattern groups ($df = 5$) and the residual variability among the means of the subgroups within the factor pattern groups ($df = 5$). The summaries of these analyses of variance can be found in Table 3. All three factor scores discriminated among the subgroups at the .01 level of confidence. The differences among the factor pattern groups were significant at the .01 level for all three scores, while the residual variation among the subgroups comprising the pattern groups did not approach statistical

TABLE 3
Analyses of Variance of Curriculum Subgroup Differences in Factor Scores

Sources of Variation	*df*	Factor A		Factor B		Factor C	
		Mean Square	*F*	Mean Square	*F*	Mean Square	*F*
Curriculum Subgroups	10	33.98	3.80**	39.69	4.67**	31.37	3.05**
Factor Patterns	5	62.09	6.95**	73.32	8.63**	52.47	5.13**
Subgroups within Patterns	5	5.86	.66	6.05	.71	10.01	.97
Within Subgroups	256	8.94		8.50		10.29	

** Significant at the .01 level.

significance. Thus all of the variation in mean factor scores could be accounted for by the grouping of the curriculum subgroups accomplished through the factor analysis of subgroup profiles reported above.

The intercorrelations among the factor scores were computed for each of the 11 subgroups and averaged (weighted mean). The average intercorrelations ($df = 245$) among the factor scores were: A and B, $r = .16$ (significant at the .05 level); A and C, $r = -.24$ (significant at the .01 level); and B and C, $r = -.05$ (not significant). Although two of the three factor score intercorrelations are statistically significant for this large sample of *S*s, all three are low enough to indicate that the scores are relatively independent measures of job incentive dimensions.

The mean factor scores for each of the pattern groups were computed and are given in Table 4. For Factor A, Group IV appears to be quite high in "need achievement" while Groups I and III show a high degree of "fear of failure." Groups V and VI have mean factor scores (Factor B) indicating strong interest in the job, while the other groups, particularly Group I, view the job as a means of advancing their own status and position. The average Factor C scores show Groups III and V to prefer jobs that are free of close and immediate contact with supervision and Group VI *S*s express a need for a strong and intimate dependency relationship with their supervisors.

TABLE 4
Mean Factor Scores of the Curriculum Factor Pattern Groups

Factor Pattern	*N*	Factor A	Factor B	Factor C
I	86	1.06	−2.17	1.06
II	42	2.69	−1.55	.76
III	46	1.87	− .80	2.63
IV	20	4.90	−1.25	1.60
V	62	2.95	.73	3.11
VI	11	2.73	1.09	.09

Discussion

From one point of view the present study had two questions: (*a*) whether the somewhat crude factor "scores" developed in the previous study (**1**) were valid in discriminating differences between subgroups of *S*s who had chosen different college curricula, and (*b*) whether a factor analytic grouping of the curriculum subgroups would account for a sizable amount of the intersubgroup variability in factor scores and provide information about the dimensions of job incentives for college *S*s. Table 1 indicates that several mistakes were originally made in defining the (assumed) homogeneous curriculum subgroups. Both the men and women college B.A. subgroups should have been further split to achieve more homogeneity; probably by separating *S*s majoring in the Humanities from those majoring in the Social Sciences. Similarly, the men in the Pre-Education subgroup should be divided into more similar subgroups. However, these errors in the a priori subgrouping of the *S*s could not be corrected after the Table 1 results were known without leaving the study open to the charge of data manipulation. These subgrouping errors appear to have had little effect on the answers to the original two questions.

The 11 subgroups condensed nicely into six broader groupings on the basis of the factor analysis reported in Table 2. Although we labelled the three obtained Factors X, Y, and Z, as "masculinity," "tenderminded social service vs. toughminded practicality," and "interest in science and technology," it must again be emphasized that we are not convinced that there are adequate appellations and that "naming" of the factors was unimportant for our purpose. We were interested solely in the use of factor analysis here to allow us to combine curriculum subgroups into a more parsimonious set of major groupings.

The answers to the two major research questions are found in Table 3. The factor scores showed differences among the 11 curriculum subgroups that were highly significant. Parenthetically it might be noted that the data given in Table 3 allows the computation of epsilon correlation coefficients to provide rough indices of the "validity" with which the Factor A, B, and C scores discriminated differences among the subgroups. Epsilon is a curvilinear correlation method similar to the eta correlation coefficient included in most statistics texts and is based upon the ratio of the variance of the means of the subgroups to the variance within the subgroups (**3,** pp. 319–324). These validity coefficients were .31, .35, and .27. The average intercorrelations among the factor scores were also low enough to estimate that each score provided some independent information about subgroup differences. Table 3 also shows that practically all of the subgroup differences in scores were accounted for by the differences among the factor pattern groups. Another way of saying this is that the curriculum subgroups included within the six larger groupings were quite homogeneous in factor scores. Curriculum differences account for approximately 10, 12, and 7 per cent of the inter-subject variability in factor scores with the six factor pattern groups accounting for practically all of this curriculum-related variance.

It seems fair to conclude from this preliminary study that (*a*) the method of factor analysis provides a technique by which the dimensions of job incentives can be isolated and relatively independent factor scores can be estimated for individual *S*s, and (*b*) these factor scores are significantly related to college *S*s' choices of academic curricula. However, further expansion and delineation of the factors underlying job incentives is needed with a consequent development of additional factor scores that can be estimated more reliably.

Summary

A list of eight job incentives was ranked by 267 college *S*s and the *S*s were divided into 11 subgroups on the basis of sex and college curriculum variables. A factor analysis of the subgroup incentive profiles isolated three factors (tentatively labelled "masculinity," "tenderminded social vs. toughminded practicality," and "interest in science and technology") and condensed the 11 subgroups into six major curriculum groupings on the basis of the subgroup patterns of factor loadings. Scores on three factors isolated in a previous factor analysis of the incentives were computed for individual *S*s and significant differences

(.01 level) were found among the six factor pattern groups for all three factor scores.

REFERENCES

1. Bendig, A. W., & Stillman, Eugenia L. Dimensions of job incentives among college students. *J. appl. Psychol.*, in Press.

2. Lyerly, S. B. The average Spearman rank correlation coefficient. *Psychometrika*, 1952, **4**, 421–428.

3. Peters, C. C., & Van Voorhis, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill, 1940.

4. Siegel, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

Received April 12, 1958.

JOURNAL OF EDUCATIONAL PSYCHOLOGY
Vol. 49, No. 5, 1958

RELATIONSHIP OF INTELLIGENCE AND SOCIAL POWER TO THE INTERPERSONAL BEHAVIOR OF CHILDREN[1]

ALVIN ZANDER AND ELMER VAN EGMOND

Research Center for Group Dynamics, University of Michigan

There are contradictory beliefs about the behavior of highly intelligent children in school, and particularly about their participation in problem-solving discussions. These children are described as both influential and impotent, tolerant and impatient, supporting and rejecting, eager and bored. Little information exists which can help us to separate fact from fancy among these assertions, nor do we know much more about the ways in which intelligent persons differ from less intelligent ones in these respects.

There is a reason for this lack of information. Intelligence as a concept is primarily intended to describe an ability to deal with cognitive problems. There are few elements in definitions of intelligence which suggest that variations in intellectual ability are associated with variations in face to face behavior. Hence, intelligence has seldom been used as an independent variable in studies of social behavior.

In a decision making group, however, a person with high intelligence may perhaps offer wiser observations than one with low intelligence. Because of his greater usefulness, we can anticipate that a brighter child would be more influential than a less intelligent one. Because he is more expert, he should have greater social power, that is, a greater ability to influence others. On the basis of assumptions like these, it is apparent that intelligence can be a cause for particular types of interpersonal behavior. Some children, regardless of their degree of intelligence are able to exercise strong influence on their peers, while others are consistently ignored by classmates.

How then does a person's intelligence affect the way he acts toward others when his group must reach a decision? Does he behave differently when he is used to having his ideas accepted (has high social power) than when he is used to being ignored? These are primary interests in this study.

Boys and girls are expected and required to behave differently in social settings. Boys, for example, are more often pressed than are girls to be concerned with achievement and influence. Because prescriptions for the two sexes differ, it is probable that the meaning and effects of intelligence or social power differ for boys and girls. A secondary purpose of this study is to examine the impact of intelligence and social power on the interpersonal behavior of boys compared to girls.

Toward these purposes, measures were made of the intelligence and social power of all children in a number of classrooms. These persons were then put in standardized, small, problem-solving groups. Their participation was observed in terms of precoded categories to see how those with different degrees of power and intelligence differed in their behavior. Data were also obtained from teachers and classmates concerning characteristics of these children in regular classrooms.

Major Issues

In the small discussion groups, we assumed children would behave in the ways

[1] The research reported herein was performed pursuant to a contract with the United States Office of Education, Department of Health, Education and Welfare.

they typically do in their classes to try to win acceptance for their ideas.

It was expected that intelligence should make a difference in the actions initiated by them, as already mentioned, because intelligent persons have more resources to offer in a problem-solving discussion than do less intelligent ones. Also, they might have more confidence in their own proposals, stemming from the ready acceptance of them in the past. Highly intelligent pupils, therefore, were expected to make more efforts to influence others, to have their ideas accepted more often than less intelligent pupils, and to behave in ways which could be taken as typical of persons with confidence in themselves.

Children with greater social power were expected to make more efforts to influence others, to be more successful in doing so, and to behave in ways which have been shown in other studies to be typical of persons with greater power **(2, 3, 4, 5, 6)**.

How do teachers characterize children who have different degrees of intelligence and social power? To determine this, teachers were asked to rate the behavior of *S*s in categories roughly similar to those used in observing the behavior of the children in the small groups. What aspects of the teachers' opinions, based on day by day experience with these children, support or contradict the behavior shown in the standardized group situations? Were teachers able to differentiate between the behavior of one type of person and that of another?

Finally, it is useful to know how classmates characterized the children to whom they ascribed high social power as compared to those to whom they attributed little power. The *Ss* were rated by their peers concerning their ability in schoolwork, their attractiveness, and their ability to coerce or threaten others. These personal qualities were considered to be separate sources of social power as suggested by French and Raven **(2)**. It was anticipated that persons with greater social power, in contrast to those with less, would have these qualities ascribed to them more often by classmates. Do highly intelligent persons differ from the less intelligent in these respects?

Method

Data used in this research were originally collected by colleagues for a different but related purpose.[2] In the original investigation, measures were obtained concerning all children in 16 second grade and 16 fifth grade classrooms, representing all socioeconomic levels in a medium-sized city. Children were selected in that study for a field experiment concerned with the creation of changes in the social position and behavior of group members. The measures were made in order to establish a baseline, so that the amount of change in a participant's behavior could be determined. The data employed in the present investigation are from these pre-experimental measures.

From the original population of 638 children on whom measurements were available, *S*s were chosen for this study who had degrees of intelligence and power in required combinations. Pupils designated as high in intelligence were those in the upper 33 per cent of their class and those designated as low in intelligence were in the lower 33 per cent. Children designated as high in power were in the upper 50 per cent of their class and those designated as low in power were in the lower 50 per cent of their class on this measure. The sample included 226 boys

[2] The collection of the original data was supported by Grant-in-Aid (M-919) from the National Institute of Mental Health, National Institutes of Health, US Public Health Service. Ronald Lippitt was project director. We are grateful for his kindness in furnishing the data for our use.

and 192 girls, 230 second grade children and 188 fifth graders.

The type of data available for each child and the source of each are as follows:

Intelligence. Intelligence scores were obtained from school records. They were based on the results of the Kuhlman-Anderson test, administered in a group form.

Ratings by classmates. Every class member rated every child in his class on four characteristics: social power (who can most often get you to do things for him?), attractiveness, ability in school work, and ability to threaten others. To obtain these ratings, photographs of every child in the classroom were printed on sheets with a four-point rating scale next to each picture.

Observed behavior in groups. The members of a class were divided into four smaller groups on a random basis and each group was sent to a corner of the room where they worked on assigned problems. A trained observer was stationed in each corner who recorded on a precoded observation schedule the quantity and quality of behavior initiated by each child.[3] Four problems were assigned to the groups, each of which required a group decision as a first step in progress toward completion of the task. In one problem, for example, the group built a large tinker-toy and in another arrived at a decision as to how many beans were in a bottle. At the end of each task new groups were formed for the next problem, thus providing maximum opportunity for each child to interact with every other child in the room. No chairmen were designated for these discussions.

The observed behavior was coded into the following categories:

1. Influence attempts—efforts to influence others regardless of the manner employed.
2. Successful influence attempts—efforts in which compliance was obtained from others.
3. Unsuccessful influence attempts—efforts in which compliance was not obtained from others.
4. Demanding influence attempts—comments made in an ordering or directing manner, implying autonomy for the actor.
5. Suggestions—comments indicating weak proposals to, or requests of, others.
6. Valuing of others—behavior indicating recognition of either high or low resourcefulness of another person in an area of knowledge or skill.
7. Positively valuing others' behavior—comments indicating recognition of high value in others' behavior.
8. Negatively valuing others' behavior—comments indicating recognition of low value in others' behavior.
9. Affect-laden behavior—behavior in which overt friendliness or unfriendliness is observed which is not a direct attempt to influence another.
10. Aggressive behavior—acts of aggression which are either inflicted or threatened.
11. Mean weighted directness in style—ratio of frequency of forceful to nonforceful forms of behavior toward others.

Perceptions by teachers. Every child was rated by his teacher on seven characteristics descriptive of the typical social behavior of the person in his schoolroom. The teacher was asked to indicate on a five-point scale the extent to which each child showed the behavior under consideration (from "hardly ever" to "most often"). The qualities rated are shown in Table 6.

Teachers did not know how much social power classmates had attributed to the *Ss* at the time they made their own ratings. Information concerning the intelligence of these children was available to teachers in school records.

Results

Characteristics attributed to the *Ss* by their peers, for boys and girls with different degrees of power, are first discussed. Results are then presented for (*a*) the observed behavior of *S*s in the small groups and (*b*) the perceptions of *Ss*' daily behavior by teachers. Finally, the results are summarized and interpreted.

For the sake of brevity the data in tables are usually limited to statistically significant findings. The omission of results for a specific category of behavior indicates that no significant differences were obtained for it.

At the outset it should be noted that

[3] The observers had satisfactory reliability among them. Information concerning reliability of the observations will be available in publications concerning the project from which these data were borrowed.

social power is not highly correlated with intelligence. The correlation for boys was .20 and for girls .28. The low relationships between intelligence and power indicates that they may vary quite independently. Since both of these correlations are significant at the .01 level of probability, however, it is also clear that brighter children tend to have more power than less intelligent ones.

Characteristics Attributed to Subjects by Classmates

Both boys and girls who were attributed high social power were more attractive to classmates than those low in power, regardless of their intelligence. These results may be seen in Table 1. Girls with greater power were rated as more able in schoolwork than girls low in power irrespective of their intelligence, while boys were described as more able in schoolwork to a significant degree only among those high in power and intelligence. Boys with higher power were described as more threatening. This was not true for girls.

Boys and girls with higher intelligence were seen by classmates as significantly more able in schoolwork than less intelligent persons (*M* 2.24 and 1.76, *p* of diff. = .01). Girls high in intelligence were rated as more attractive than those low in intelligence (*M* 2.40 and 2.01, *p* of diff. = .005). Ability to threaten was not related significantly to intelligence.

Behavior of Boys in Small Groups

We consider the types of behavior boys used in the problem solving discussions.

Effects of Variations in Power

Among highly intelligent boys, those high in social power were not significantly different in any category of observed behavior from those low in power.

Among less intelligent boys, those high in power, compared to those low in power, revealed a vigorous, inconsistent, and com-

TABLE 1
Means of Characteristics Attributed by Classmates to Children with Varied Social Power

Children high in intelligence	Boys			Girls		
	Social Power		*t*	Social Power		*t*
	High *M*	Low *M*		High *M*	Low *M*	
Attractive	2.13	1.52	6.42**	2.09	1.50	7.66**
Able in school work	2.19	1.43	6.33**	2.14	1.58	7.18**
Threatening	3.02	2.76	3.29**	2.89	2.86	.39
N	58	38		73	42	
Children low in intelligence						
Attractive	2.21	1.59	6.59**	2.39	1.65	11.69**
Able in school work	2.44	1.73	1.15	2.55	1.76	16.28**
Threatening	2.91	2.78	1.86*	2.98	2.90	.90
N	46	84		26	51	

* $p = .05$.
** $p = .001$.

petitive pattern of behavior. The quantitative results may be seen in Table 2.

Effects of Variations in Intelligence

Among boys with high power, the behavior of the more intelligent was in no way significantly different from that shown by the less intelligent.

Among boys with low power, those with greater intelligence, compared to those with less, were active, effective, and supportive of others. These data are summarized in Table 3.

Behavior of Girls in Small Groups

Effects of Variations in Power

Among girls with high intelligence, those with high power were not observed to behave differently from those with low power.

Among girls who were less intelligent, those with high social power were only a little different from ones low in power. Those with more power were more often successful in their influence attempts (*M* 8.81 and 5.61, *p* of diff. = .01) and were more positive in their remarks about others' behavior (*M* 1.28 and .58, *p* of diff. = .01).

Effects of Variations in Intelligence

Regardless of their social power, variations in intelligence were not associated to a significant degree with differences in girls' observed behavior.

TABLE 2

Behavior of Less Intelligent Boys Varying in Social Power

	Social Power		*t*
	High *M*	Low *M*	
Total influence attempts	24.34	16.00	3.18**
Freq. successful influence attempts	13.87	7.96	3.16**
Freq. unsuccessful influence attempts	10.48	8.04	1.98*
Freq. demanding influence attempts	5.33	3.46	2.10*
Freq. suggestions	18.90	12.30	3.13**
Total valuing of others' behavior	2.48	1.73	2.03*
N	46	84	

* $p = .05$.
** $p = .001$.

TABLE 3

Group Behavior of Boys with Low Social Power Varying in Intelligence

	Intelligence		*t*
	High *M*	Low *M*	
Total influence attempts	21.63	16.00	2.02*
Freq. successful influence attempts	11.97	7.96	2.06*
Freq. suggestions	16.86	12.30	2.04*
Freq. positively valuing others' behavior	1.37	.89	2.58**
N	38	84	

* $p = .05$.
** $p = .01$.

Behavior of Boys and Girls Directly Compared

A direct comparison of the behavior used by boys and girls, regardless of their power or intelligence indicates the actions most typical of each sex. The following behaviors were observed among boys significantly more often than among girls: attempts to influence, successful influences, unsuccessful influences, aggression, and demands. Girls did not display any type of behavior significantly more often than boys. It is evident that boys were considerably more active and demanding in their groups than were girls.

TABLE 4

Behavior of Highly Intelligent Boys Versus Girls Varying in Social Power

	Social Power			Social Power		
	Low Boys *M*	Low Girls *M*	*t*	High Boys *M*	High Girls *M*	*t*
Total influence attempts	21.63	13.57	2.30*	25.55	16.69	3.08***
Freq. successful influence attempts	11.97	7.21	1.80*	15.76	9.79	2.62***
Freq. unsuccessful influence attempts	9.66	6.36	2.39**	9.78	6.90	1.79*
Freq. suggestions	6.87	10.04	3.13***	19.95	13.80	2.75***
Total valuing of others' behavior	2.16	1.67	.97	2.29	2.09	.48
Freq. demanding influence attempts	4.63	3.14	.28	5.17	2.61	2.64***
Freq. aggressive behavior	2.97	1.50	1.89*	2.79	1.54	2.36**
Total affect-laden remarks	1.95	2.02	.15	2.46	1.63	3.92***
M weighted demandingness in behav.	21.15	20.95	.31	22.64	21.25	1.50
N	38	42		58	73	

* $p = .05$.
** $p = .01$.
*** $p = .001$.

When members of both sexes were high in intelligence but low in power, boys were more active and aggressive than were girls. These data are shown in Table 4.

Where members of both sexes were high in both intelligence and power, the patterns of behavior were similar to those just described. In addition, high-high boys were demanding in their comments and used more affect-laden types of behavior than high-high girls (see Table 4).

In contrast to the girls, then, highly intelligent boys were likely to be active in their groups regardless of their social power and likely to be aggressively insistent in stating their opinions when high in both power and intelligence.

Among the less intelligent children the boys again appeared to be more involved in their groups than the girls if they were high in power. A comparison of the behavior of boys and girls is presented in Table 5. Boys with low intelligence and low power were very little different from girls with low intelligence and power.

Teachers' Perceptions of Boys

We turn to an examination of the qualities teachers attributed to these children.

Effects of Variations in Power

Among highly intelligent boys, the teachers made clear distinctions on every category between boys who were high in power and those who were low in power. These results are shown in Table 6.

Among less intelligent boys, the teachers made similar distinctions in the characteristics they attributed to boys high in power and those low in power. These re-

TABLE 5

BEHAVIOR OF LESS INTELLIGENT BOYS VERSUS GIRLS AS RELATED TO SOCIAL POWER

	Social Power			Social Power		
	High Boys *M*	High Girls *M*	*t*	Low Boys *M*	Low Girls *M*	*t*
Total influence attempts	24.34	15.96	2.44**	16.00	13.49	1.09
Freq. successful influence attempts	13.87	8.81	1.99*	7.96	5.61	1.59
Freq. unsuccessful influence attempts	10.48	7.15	1.90*	8.04	7.88	.12
Freq. suggestions	8.89	13.23	2.07***	12.31	9.75	.84
Freq. positively valuing others' behavior	1.19	.61	2.04*	.89	.58	1.63*
Total valuing others' behavior	2.48	1.54	1.96*	1.76	1.41	.99
M. weighted demandingness	21.98	19.70	.30	21.95	20.09	2.86***
Freq. aggressive behavior	3.24	2.00	1.40	3.13	2.20	1.50
N	46	26		84	51	

* $p = .05$.
** $p = .01$.
*** $p = .001$.

sults are listed in Table 6. In the eyes of teachers, boys with greater power are strikingly different in their social behavior from those with less power.

Effects of Variations in Intelligence

Teachers made no distinctions to a significant degree between boys high in intelligence and those low in intelligence, regardless of their social power.

TEACHERS' PERCEPTIONS OF GIRLS

Effects of Variation in Power

Among girls high in intelligence, those with high social power were seen to be different from girls low in power as shown in Table 7.

Among girls low in intelligence those with high social power were perceived by teachers to be different from those with low social power in only two respects. Girls with higher power were seen as more often successful in influencing others (*M* 4.82 and 3.77, *p* of diff. = .005) and more friendly (*M* 4.30 and 3.46, *p* of diff. = .005).

Effects of Variations in Intelligence

Among girls with greater social power, teachers saw girls with high intelligence as being little different in their classrooms from those with low intelligence: the girls were viewed as making more attempts to influence classmates (*M* 4.26 and 3.80, *p* of diff. = .02) and as more successful in these attempts (*M* 3.77 and 3.19, *p* of diff. = .005).

Among girls with little social power, highly intelligent girls compared to less intelligent ones were seen by teachers as making more frequent attempts to exercise influence in the class (*M* 4.29 and 3.76, *p* of diff. = .02), as more successful in these

TABLE 6

Characteristics Teachers Attribute to Boys with Different Degrees of Social Power

	Boys high in intelligence			Boys low in intelligence		
	Social Power			Social Power		
	High *M*	Low *M*	*t*	High *M*	Low *M*	*t*
Freq. influence attempts	4.05	3.70	1.50	4.10	3.52	2.40**
Amt. success as influencer	4.10	3.24	3.69***	4.46	3.50	4.38***
Friendliness	3.95	3.41	2.43**	4.27	3.41	3.63***
Depends on teacher	3.57	4.62	3.84***	3.76	4.32	3.08***
Depends on peers	3.68	4.45	3.00***	3.56	4.04	2.23*
Self-centeredness	3.58	4.26	2.93***	4.03	3.68	1.60
Degree of forcefulness	3.68	4.44	3.00***	3.56	4.04	2.23*
N	58	38		46	84	

* $p = .05$.
** $p = .01$.
*** $p = .001$.

efforts (*M* 4.82 and 3.78, *p* of diff. = .005), and as more friendly (*M* 4.29 and 3.62, *p* of diff. = .005).

Teachers' Perceptions of Boys and Girls Compared

The perceptions teachers had of boys and girls differed to a significant degree only where the members of both sexes were high in power while low in intelligence. Here, boys were seen as more active in making attempts to influence classmates and more self-centered in doing so than were girls.

Differences Due to Ages of Subjects

The results on all dimensions for second graders were compared with those of fifth graders. There was no acceptable evidence that differences in ratings of *S*s by peers or teachers, or in their observed behavior in the small groups, were associated with differences in age.

TABLE 7

Characteristics Teachers Attribute to Highly Intelligent Girls with Different Degrees of Social Power

	Social Power		*t*
	High *M*	Low *M*	
Amt. success as influencer	3.78	3.19	3.39**
Friendliness	3.62	3.07	2.75**
Depends on teacher	4.00	4.38	1.80*
Self-centeredness	3.93	4.63	3.16**
N	73	42	

* $p = .05$.
** $p = .001$.

Summary and Discussion of Results

Characteristics Attributed by Peers

Boys and girls were similar in that those with greater power, compared to those with less, were better liked by their peers.

Attractiveness then was an important basis of power for both sexes.

Boys and girls were different in that boys with greater power were seen by classmates as more threatening, and girls with greater power were rated as more expert "in the things you do in school."

These results suggest that among these children the two sexes won social power in part on the basis of different attributes, boys earned it by being threatening and girls earned it by being skilled in the things required of a school child.

Observed Behavior in Groups

The behavior of boys may be briefly summarized by noting that those who were low in social power and low in intelligence (low-lows) were strikingly different from boys with all other combinations of intelligence and power. The significant differences we have seen in the behaviors of boys stem primarily from the fact that the low-lows were passive persons in their groups, more than boys who were high in either power or intelligence. The low-lows significantly less often tried to influence others, were less successful in doing so, were less demanding in manner, less often evaluating in respect to others' contributions, and less often suggestors of tentative proposals. Boys with various combinations of power and intelligence other than low-low were in no way different from one another to a statistically significant degree in their observed behavior, that is, high-highs were not different from high-lows or from low-highs. It is worth special note that highly intelligent boys whose social power was low behaved in no way different from those whose social power was high.

Because the low-lows were different from boys with more power or intelligence in ways that were quite comparable, and because the high-highs were in no way different from the high-lows or low-highs; it appears that the possession of greater intelligence may be the same as the possession of greater power insofar as the effect upon behavior in these groups is concerned.

This similarity is due, we believe, to the fact that the contributions made by highly intelligent boys in a problem solving discussion represent resources which are valuable to the group. The possession of these resources, we assume, provided power for the owners to influence those who valued them. It is highly probable, although it cannot be demonstrated with the data available here, that boys with high intelligence were treated by group members as though they were persons with high power. As a consequence, boys with high intelligence became aware of the value of their ideas and of the influence they were having on the discussion. Although they had little power accorded to them previously by their classmates, boys with high intelligence apparently behaved in the groups like those who had come to the group with social power already attributed to them by their peers.

It is noteworthy that boys low in intelligence yet high in social power tended to be more inconsistent than low-lows in that they made both demands and weak proposals, and praised as well as criticized others, whereas boys high in intelligence but low in power more consistently proposed ideas and supported others' behavior than did the low-lows. The possibility is suggested by these findings that greater power among boys generated more inconsistent and coercive group behavior while greater intelligence evoked a consistency and readiness to be considerate in relations with others. This conjecture is supported by the findings that peers characterized boys with greater power as threatening but did not so describe boys with greater intelligence.

Girls who were low-lows were less successful in influencing their groups and less often made positive remarks than girls high in power but low in intelligence. This indicates that low-low girls, like the low-low boys, were passive in their groups' discussions. Low-low girls were not different in any respect from girls low in power and high in intelligence, which suggests that high intelligence was not similar to power among girls, as was earlier noted for boys.

In most combinations of intelligence and power, boys were significantly more active in seeking to influence others, more often successful, more often unsuccessful, and more likely to evaluate the comments made by others than were girls. Low-low boys were very much like girls and most like low-low girls, differing from them in only two respects: they were significantly more likely to be demanding and to be positive in commenting upon the contributions made by others.

In sum, the possession of either power or intelligence by boys appeared to stimulate vigorous and successful participation in their groups' work while the possession of low power and low intelligence together generated passivity in boys. The amount of intelligence or power a young person possesses affected a boy's behavior more than it did a girl's behavior in these problem solving discussions.

To explain the consequences of power and intelligence for boys and girls we assume that these two attributes have a differential significance for the sexes in allowing them to conform to the expectations society puts upon them. Barry, Bacon, and Child (**1**) have reported that boys are expected to be self-reliant and to strive for achievements, while girls are urged to be nurturant, obedient, and responsible in almost all societies, including ours.

We assume that, in their group behavior, the boys and girls were attempting to conform to these prescriptions for their sexes. Boys who were high in either intelligence or social power more clearly fulfilled their sex roles than those lacking in both of these. Boys who were low in both intelligence and power least often showed the behavior required of their sex. The low degree of their intelligence and power apparently made them unable to perform in ways typically expected of them. Thus, either social power or intelligence was necessary for boys if they were to act as boys are expected to act. There is some indication that social power was more important than intelligence in this respect.

Girls who were high in power and intelligence were little different from those who were low in either of these qualities because, we believe, high social power and intelligence were not needed in order to be the nurturant, obedient, or responsible persons required by society. Girls could fulfill these expectations regardless of the amount of power or intelligence they possessed.

Perceptions by Teachers

The children who were accorded higher social power by their classmates were viewed by teachers as most influential, since the teachers rated boys and girls with higher power as more successful in influencing others than those low in social power.

On almost every category, teachers distinguished between boys low in power and those high in power, regardless of the boys' intellectual abilities. It is striking in this connection that teachers saw those with greater power as less forceful and more friendly than those with less power, whereas boys with greater power (and low intelligence) were demanding and less friendly than those with less power when observed in the small groups. Clearly teachers did not perceive the boys with greater social power as threatening persons in the way that classmates saw them.

In their considerations of girls, apparently those who were high in either intelligence or power were seen as more effective members of their classes than those low in these respects.

Why did the teachers characterize boys who had greater power (but low intelligence) as more considerate of others than was observed among them in the small groups? The most likely explanation is that their behavior actually was different in their classrooms from what it was in the discussions. In a class the teacher is in charge so that nondemanding, friendly behavior is required and becomes the standard way for interacting with peers during school. It is also probable that teachers approved of achievement oriented and influential efforts when they were used by boys since such behavior is in accord with the demands that teachers, among other adults, place upon young males. Thus, teachers attributed positive characteristics to the boys whom they viewed as most influential in their classrooms.

In the small groups no one was in charge. Hence, the boys with greater power, but less intelligence, were free to insist upon having their opinions accepted and were free to use coercion when necessary in order to be influential.

Teachers did not perceive differences in classroom behavior between boys who were high in intelligence and those who were low in intelligence although differences between these two groups were evident in the problem solving discussions. It appears that variations in intelligence do not generate variations in social behavior in a classroom, of the kind that teachers were asked to report.

We assume that girls were less active in their classes and therefore less visible to the teachers than the boys. Thus, teachers made fewer distinctions in characterizing the behavior among girls with different amounts of power than they did among boys. Girls with higher power were seen by teachers as most friendly to classmates, which suggests that through this friendliness they won influence in the classroom. Girls with higher intelligence were also seen by teachers as being more friendly and as making more attempts to influence others than girls with low intelligence. It is probable that the girls with high intelligence did use acceptable forms of social influence in their schoolroom relations since they were rated as attractive by their classmates, more so than those with low intelligence. This acceptance by their peers apparently gave them confidence to exercise their influence freely and teachers noted this in characterizing highly intelligent girls.

In summary, teachers perceived the behavior of girls as pretty much alike regardless of their power or intelligence. They made no significant distinctions among boys with different degrees of intelligence, but they saw many distinctions in the behavior of boys who differed in social power. Assuming that the teachers' perceptions are accurate, it is evident that a boy's social power determines his behavior in the classroom more than his intelligence does, whereas differences in the power or intelligence of a girl has little effect upon the behavior she employs in the schoolroom.

REFERENCES

1. Barry, H., Bacon, Margaret, & Child, I. A cross-cultural survey of some sex differences in socialization. *J. abnorm. soc. Psychol.*, 1957, **55**, 327–332.
2. French, J. R. P., Jr., & Raven, B. The bases of social power. In D. Cartwright (Ed.), *Studies in social power*. Ann Arbor, Mich.: Institute for Social Research, in press.
3. Hurwitz, J., Zander, A., & Hymovitch, B. Some effects of power on the relations

among group members. In D. Cartwright and A. Zander (Ed.), *Group dynamics research and theory*. Evanston, Ill.; Row, Peterson, 1953, 483–492.

4. Lippitt, R., Polansky, N., Redl, F., & Rosen, S. The dynamics of power. *Hum. Relat.*, 1952, **5,** 37–64.
5. Zander, A. & Cohen, A. R. Attributed social power and group acceptance. *J. abnorm. soc. Psychol.*, 1955, **51,** 490–492.
6. Zander, A., Cohen, A. R., & Stotland, E. *Role relations in the mental health professions*. Ann Arbor, Mich.: Institute for Social Research, 1957.

Received June 4, 1958.

Journal of Educational Psychology
Vol. 49, No. 5, 1958

AN APPROACH TO THE MEASUREMENT OF PSYCHOLOGICAL CHARACTERISTICS OF COLLEGE ENVIRONMENTS[1]

C. ROBERT PACE AND GEORGE G. STERN
Syracuse University

The present article considers the idea that college cultures may be seen as a complex of environmental press which, in turn, may be related to a corresponding complex of personal needs. In the psychological literature, one is indebted to Henry Murray (**5**) for the dual concept of personal needs and environmental press. In the broadest sense, the term "need" refers to denotable characteristics of individuals, including drives, motives, goals, etc. The term "press" can similarly be regarded as a general label for stimulus, treatment, or process variables. Murray's concept of needs has provided a starting point for the construction of various objective measures of personality (**2, 3, 4**). No parallel development in the objective measurement of environmental press, however, has previously been attempted. College students differ. College environments also differ. The concept of press offers a way of viewing the environment which is comparable analytically and synthetically to the more familiar ways of dealing with the individual.

Development of the College Characteristics Index

Using Murray's classification of needs as a model, Stern has constructed several experimental editions of a needs inventory, called the Activities Index. In its current form the Activities Index consists of 300 statements of commonplace, socially acceptable activities to which responses of "like-dislike" are given. There are 30 scales of 10 items each, corresponding to 30 needs in Murray's taxonomy. Some scales can be scored positively or negatively, as for example conjunctivity-disjunctivity, succorance-autonomy, impulsion-deliberation, etc., so that the total number of needs to which scores can be attached is 42 rather than 30. A preliminary manual for the Activities Index describes the test in detail (**8**).

A corresponding test for describing college environments, called the College Characteristics Index, was subsequently constructed. It consists of 300 statements about college environments to which responses of "True-False" are given. The statements are organized into 30 ten-item scales, with a press scale for each need scale that was included in the Activities Index. The following kind of questions guided the writing of items: what might be characteristic of an environment which exerted a press toward order, or toward autonomy, or toward nurturance, or understanding, or play, etc? Stated in another way, what might there be in a college environment which would be satisfying to or tend to reinforce or reward an individual who had a high need for order, or autonomy, or nurturance, or understanding, or play, etc.? The items themselves are statements about college life. They refer to the curriculum, to college teaching and classroom activities, to rules and regulations and policies, to student organizations and activities and interests, to features of the campus, etc.

Sample items from corresponding Need and Press scales will illustrate the parallelism.

A need for Order would be inferred from liking such activities as: "Arranging my clothes neatly before going to bed. Hav-

[1] Portions of this research have been financed by grants from the College Entrance Examination Board and the Carnegie Corporation.

ing a special place for everything and seeing that each thing is in its place. Keeping a calendar or notebook of the things I have done or plan to do." What might such a person like to find in a college environment or what features of a college environment might be rewarding or frustrating to such a need? The following items from the press scale for Order might be relevant: "Faculty members and administration have definite and clearly posted office hours. In many classes students have an assigned seat. Professors usually take attendance in class."

On the need scale for Impulsion-Deliberation, a high score for Impulsion results from liking such activities as: "Being in a situation that requires quick decisions and actions. Doing things on the spur of the moment. Doing whatever I'm in the mood to do." Thus, a college environment which has a press toward impulsiveness might be a place where: "Most students don't decide what courses to take until the time of registration. Students often start projects without trying to decide in advance how they will develop or where they may end. Spontaneous student rallies and demonstrations occur frequently."

A high need for Energy is inferred from liking such activities as: "Taking up a very active sport. Having something to do every minute of the day. Giving all of my energy to whatever I happen to be doing." The needs of such a person might be expected to find fulfillment and satisfaction in a college environment where: "There is an extensive program of intramural sports and informal athletic activities. Student gathering places are typically active and noisy. Class discussions are typically vigorous and intense."

Just as needs are inferred from the characteristic modes of response of an individual, so press are reflected in the characteristic pressures, stresses, rewards, conformity-demanding influences of the college culture. Operationally, press are the characteristic demands or features as perceived by those who live in the particular environment. To each statement in the College Characteristics Index the person who takes the test answers true if he believes it is generally characteristic of the college, is something which occurs or might occur, is the way people tend to feel or act; and he answers false if he believes it is not characterisic of the college, is something which is not likely to occur, is not the way people typically feel or act.

A Study of Five Institutions

A first draft of the College Characteristics Index was administered in May, 1957, to groups of students at five institutions and to smaller groups of faculty members at four of the five institutions. In all, 423 students and 71 faculty members responded to the instrument. Neither student nor faculty groups were representative samples. Most of the students were upperclassmen and most of the faculty members were in the upper academic ranks. It was argued that if a dominant press really exists in a particular environment almost any group of people living in the environment would probably identify it. The testing program was, in any case, intended only as a preliminary try-out of the model from which some information would be gained about the types of items, the possible reliability and validity of the scales, and the potential utility of this approach to measuring college characteristics.

The five institutions, although not identified here, were selected because observers would probably agree that they are rather different from one another, with the selection of colleges thus providing some evidence about the construct validity of the test. One was a large Midwestern state university. The second was a large Midwestern private university. The third institution was a large Eastern private university. The fourth was a moderate-sized Eastern private college for men. The fifth institution was a publicly supported college in the metropolitan New York area.

Test Results

Saying that a particular press is or is not characteristic of an institution is an arbitrary matter. There exist no conventions or experience to guide the decision. The basis for the tentative decisions that were made, together with other statistical information about the scales and the items are given in the following paragraphs.

In examining means, or a profile on which mean scores are plotted, one naturally looks for what appear to be high and low points. And in examining variances of distributions, in the context of press identification, one naturally looks for some concentration of scores or a low variance, suggesting a consistency of impression, rather than a wide dispersal of scores, suggesting divergent impressions about the press.

Table 1 shows the means and standard deviations on each of the press scales, computed from the students' responses at the five institutions studied. Each of these 30 scales has a maximum possible range of 0 to 10. The median of these mean scores is approximately 5.5. The median of the standard deviations is approximately 1.7. Putting this information together, one might suggest that, for the five institutions represented, a fairly reasonable definition of a noticeable press (or its absence) would be one which required a mean score falling in the upper or lower one fourth of the total distribution. Mean scores of 6.6 or higher and mean scores of 4.4 or lower would thus be suggestive of a press. In Table 1, means which meet this criterion are in italics.

A corresponding table showing faculty means and sigmas is not presented, mainly because the number of cases is quite small—*N*s being 25, 20, 11, and 15. Also, from an educational view, it may be argued that the effective press of an environment is what the students say it is, not what the faculty say, or what the catalogue says, etc. Nevertheless, in the four colleges where both student and faculty responses to the CCI were obtained it is of some pertinence to note their similarity. Table 2 shows a distribution of the differences between the means of students' and faculty responses in each institution and also the differences between faculty and student responses to individual items. Here it is evident that most of the differences are fairly small. In nearly half of all the comparisons between faculty and student mean scores the difference between the two was less than half a point. Three fourths of all the differences were less than 1.00. Differences between faculty and student responses to individual items are grouped in three arbitrary categories. Thus, 44% of all the items answered by students and faculty were answered within a range of agreement of 10 percentage points or less; and on 12% of the items the percentage for the students' answer differed by 30 points or more from the percentage for the faculty's answer. Considering the number of cases in all groups, one can be confident that differences of 10 points or less are in all instances merely chance differences, and that differences of 30 points or greater may in most instances be significant differences at least at the 5% level of confidence. It appears then, that some of the differences in the middle category of the table are chance differences and some are significant differences. One might estimate that, over all, about three fourths of the items were answered in tolerably good agreement by both students and faculty, and that perhaps as many as one fourth of the item responses represent divergent views between students and faculty in characterizing the institution.

TABLE 1

Means and Standard Deviations on the Various Press Scales from Students' Responses at Five Institutions

Press Scales	College A (N = 100)		College B (N = 44)		College C (N = 100)		College D (N = 68)		College E (N = 111)	
	$\overline{M}$	*SD*	$\overline{M}$	*SD*	$\overline{M}$	*SD*	$\overline{M}$	*SD*	$\overline{M}$	*SD*
Abasement	*4.0*	2.2	*3.6*	1.7	*3.3*	1.9	*3.4*	1.7	*4.3*	2.1
Achievement	4.9	1.9	5.3	1.4	*4.4*	1.9	5.4	1.9	5.5	1.8
Adaptiveness	6.3	1.8	*7.9*	1.4	*6.9*	1.5	6.3	1.5	5.3	1.6
Affiliation	5.4	2.0	5.6	2.2	5.8	1.7	*4.2*	2.1	*3.8*	2.0
Agression-Blameavoidance	*4.4*	1.9	5.8	1.4	5.6	1.6	4.7	1.6	*3.5*	2.0
Change-Sameness	5.6	1.3	*7.1*	1.4	5.1	1.7	5.1	1.4	4.8	1.4
Conjunctivity-Disjunctivity	*6.7*	2.0	5.8	2.3	*7.4*	1.7	5.8	2.1	*6.9*	1.9
Counteraction	6.1	1.4	6.5	1.4	5.0	1.5	5.4	1.6	5.9	1.4
Deference	5.6	1.7	5.1	1.4	5.6	1.7	5.4	1.4	5.5	1.5
Dominance	5.5	1.5	*3.5*	1.5	4.9	1.7	5.7	1.6	*4.3*	1.8
Ego-Achievement[a]	*6.9*	2.0	5.2	1.5	5.6	1.7	*6.9*	1.8	4.9	2.2
Emotionality-Placidity	6.3	1.5	5.8	1.6	6.3	1.5	5.9	1.9	5.3	1.8
Energy-Passivity	5.1	1.9	6.0	1.7	*4.4*	1.8	5.0	1.6	4.8	1.7
Exhibitionism	*6.7*	1.7	5.8	1.6	6.2	1.4	6.3	1.5	5.6	1.8
Fantasied Achievement[b]	6.0	1.8	6.5	1.6	5.7	1.6	4.6	1.7	5.7	1.6
Harmavoidance	4.7	1.5	4.8	.9	*3.4*	1.3	*3.9*	1.3	5.5	1.5
Humanism[c]	5.7	1.9	*8.5*	1.3	6.0	1.8	*6.8*	1.8	*6.6*	2.0
Impulsion-Deliberation	*3.5*	1.5	*4.3*	1.5	*4.1*	1.5	*4.2*	1.3	*3.3*	1.4
Narcissism	5.8	1.5	*3.2*	1.3	*3.8*	1.6	5.3	1.7	5.5	1.7
Nurturance	*7.4*	1.6	5.8	1.8	*6.7*	1.7	*6.9*	1.7	6.5	1.7
Objectivity-Projectivity	*6.6*	1.9	*8.1*	1.4	*7.4*	1.9	*6.7*	1.9	6.4	2.1
Order	*7.3*	1.5	*3.3*	1.5	5.8	1.4	6.2	1.4	*7.1*	1.3
Play	*7.4*	1.5	*3.8*	1.6	6.1	1.4	*7.8*	1.6	5.9	1.9
Pragmatism[d]	4.6	1.6	*1.7*	1.1	4.9	1.6	4.7	1.5	*3.8*	1.6
Reflectiveness[e]	5.5	1.9	*8.0*	1.4	4.9	1.9	5.5	1.9	5.6	1.9
Scientism[f]	6.3	1.9	*8.4*	1.5	5.1	2.0	6.5	2.0	6.5	1.7
Sentience	5.3	1.8	6.5	1.5	*3.5*	1.6	5.2	2.0	5.0	1.8
Sex	6.2	1.4	4.9	1.5	5.7	1.3	*7.0*	1.3	4.7	1.3
Succorance-Autonomy	5.8	1.3	4.6	1.6	5.2	1.4	5.1	1.9	5.1	1.6
Understanding	4.8	1.8	*8.6*	1.2	5.1	1.7	4.7	1.6	5.8	1.8

[a] Derived from Exocathection-Intraception.

[b] Derived from Ego Ideal.

[c] Derived from Endocathection-Extraception: Social Sciences and Humanities.

[d] Derived from Exocathection-Extraception.

[e] Derived from Endocathection-Intraception.

[f] Derived from Endocathection-Extraception: Natural Sciences.

Reliability

Test-retest data are not available in the present study. Conventional estimates of reliability from a single administration may not be appropriate for an instrument on which one hopes to find skewed distributions and minimal dispersion of scores. Faculty-student agreement within the same institution is of some relevance but one might argue that the perceptions of these two groups could differ on individual items or scales, yet each could be reliable within itself.

Item analysis data are more directly relevant to the reliability of scales. An item analysis was made of each scale, separately for each of the five institutions, using the students' responses. Ebel's (1) simplified method was used. Each of the 30 scales was thus subjected to item analysis in five different samples, and since there are 10 items in each scale, the total number of item discrimination indexes obtained was 1500. Of these 1500 discrimination indexes, 1% were negative, 18% fell between .0 and .19, 30% fell between .20 and .39, and 51% were .40 or higher. In other words, 81% of the items had, on the average, moderate to high discrimination in their respective scales.

Perhaps the most important approach is one which treats reliability and validity as inseparable and deals with the instrument as a whole. For example, do different people characterize the institution in the same way? This involves the reliability of profiles, with all their interrelationships. As a first approximation of this, the rank order of mean scores from the students' responses can be compared with the rank order of mean scores from faculty responses within the same institution. Thus, do these groups see the institution in relatively the same pattern? For the two colleges (Colleges B and C) which had the largest number of faculty respondents, these rank order correlations were .96 and .88. Correlations were not computed for the other two colleges where faculty mean scores would be based on *N*s of 15 and 11.

Validity

To illustrate the sort of interpretation and description of a college environment which can be derived from the CCI, the press of two colleges are presented. The comments are based entirely on the arbitrary definitions of what levels of scores constitute a press (which was explained earlier); and the nature of these press is further illustrated by citing some of the specific items which most clearly define them. No estimate is presently available of the validity of these descriptions against a systematic outside criterion. But the descriptions show quite clearly that these are very different environments, and that the test is therefore capable of revealing some sharp distinctions between colleges which qualified observers would expect to be different. The evidence is therefore relevant to the property of validity. Detailed data, and descriptions of the other three colleges, may be found elsewhere (7).

TABLE 2

SUMMARY OF DIFFERENCES BETWEEN FAÇULTY AND STUDENT RESPONSES WITHIN EACH OF FOUR INSTITUTIONS

Mean Differences in Scale Scores	Number of Differences	Cumulative Per Cent
.0 to ±.49	57	48%
±.50 to ±.99	34	76%
±1.00 to ±1.49	23	95%
±1.50 to ±1.99	6	100%
	120	
Percentage Differences in Responses to Individual Items		
10 percentage points or less	528	44%
Between 11 and 29 points	528	88%
30 points or more	144	100%
	1200	

College A.

The major press in College A are toward orderliness and friendly helpfulness, with overtones of spirited social activity. This is suggested by high scores on the scales for Order, Objectivity, Conjunctivity, Nurturance, Play, Ego Achievement, Exhibitionism, and by low scores on the scales for Abasement, Impulsion, and Aggression.

The stress on Order, Deliberation (opposite of Impulsion), and Conjunctivity is indicated by such highly shared observations as the following: students have assigned seats in some classes, professors often take attendance, papers and reports must be neat, buildings are clearly marked, students plan their programs with an adviser and select their courses before registration, courses proceed systematically, it is easy to take clear notes, student activities are organized and planned ahead.

Within this orderliness, student life is spirited and a center of interest. For example, big college events draw lots of enthusiasm, parties are colorful and lively, there is lots to do besides going to classes and studying, students spend a lot of time in snack bars and in one another's rooms, and when students run a project everyone knows about it.

At the same time, amid this student-centered culture, there is a stress on idealism and service. Students are expected to develop an awareness of their role in social and political life, be effective citizens, understand the problems of less privileged people, be interested in charities, etc.

The total picture of the environment, then, is one of high social activity, esprit de corps, and enthusiasm combined with an emphasis on helping others and idealistic social action and all within a fairly well understood set of rules and expectations which are deliberative and orderly. One would expect some of the explicit objectives of such an institution to stress personal and social development, idealism and social action, and civic responsibility.

College B.

Here the dominant press of the environment fall in the theoretical-intellectual category—Reflectiveness, Humanism, Scientism, Understanding, and Objectivity. This dominant press occurs in an environment also characterized by Change, nondefensive acceptance of criticism (Adaptiveness), and by resistance to any abject acceptance of criticism or presumed low status (Abasement). Moreover, on two of the scales which defined a high press at College A (Play and Order), the press at B is exactly opposite. There is further, a minimum of importance to social status or manipulation for tangible ends (Pragmatism), preoccupation with self and personal appearance (Narcissism), and bossing or directing others (Dominance). There is, however, a generally consistent and high press toward Deliberation or planning and thinking ahead.

It is clear that the most pervasive press is directed toward the pursuit of understanding for its own sake, abstract and unencumbered by requirements for practical utility or social action.

The theoretical-intellectual press of the environment at College B is more specifically suggested by the following observations with which, generally, more than nine tenths of both faculty and students agree: there are excellent library resources in natural science and social science, a lecture by an outstanding philosopher or scientist would draw a capacity audience, many students are planning graduate work or careers in science or social science, there are many opportunities for students to see and hear and criticize modern art and music, reasoning and logic are valued highly in student reports and discussions, students who spend a lot of time in a science laboratory or in trying to analyze or classify art or music or in seeking to develop a personal system of values are not regarded as odd, scholarship and intellectual skills are regarded as more important than social poise and adjustment, there is time for private thought and reflection, one need not be afraid of expressing extreme views, the faculty and administration are tolerant and understanding in interpreting regulations.

In contrast with College A, students at B do not have an assigned seat in class, professors do not take attendance, students are likely to study over the weekend, big college events draw no great enthusiasm, and the place is not described as one where "everyone has a lot of fun."

Moreover, student leaders have no special privileges, family status is not important, students are not much concerned about personal appearance and grooming, and an intellectual is not an "egghead."

And finally, exams are not based on factual material from a textbook, classes are not characterized by recitation and drills, grade lists are not publicly posted, students are not publicly reprimanded for mistakes, student organizations are not closely supervised, students tend to stay up late at night, work all the harder if they have received a low grade, and if confronted with a regulation they do not like they will try to get it changed.

One would expect the explicit objectives of such an institution to stress the acquisition of knowledge and theory, critical judgment and independence, and a sense of the significance of intellectual life.

Other Differences Between Colleges

That College A and College B are quite different environments can be suggested statistically as well as in the descriptive profiles just presented. The rank order of mean scores for College A correlates, for example, .06 with the rank order of mean scores for College B. One can also estimate, without computing, the significance of the differences between the mean scores of College A and College B on each of the 30 scales. Given *N*s of 100 and 44, and standard deviations as large as 2.00, a difference of 1.00 is significant well beyond the 1% level of confidence. On at least 22 of the 30 scales, the differences between means for the two schools are significant. Considering all five schools, it is clear that on every scale except one (Deference) there are significant differences between two or more of the mean scores.

Another indication of differences can be obtained from noting the differences between the percentages of students at Colleges A and B answering each item according to the key. In the middle range of percentages a difference of 20 points, with *N*s of 100 and 44, is always significant beyond the 1% level. On this conservative basis there were 172 of the 300 items on which the percentages for College A differed significantly from College B.

The point to these observations is merely to suggest that the first trial run of the CCI produced many results which clearly differentiated among the environments or press of the five colleges. The actual items in the CCI are not reproduced here because there seems little virtue in printing a first draft. The content of many items has been, of course, indicated in the descriptive profiles of the two colleges and in the earlier paragraphs which noted the parallel structure of Need items and Press items.

Conclusions and Implications

After completing the preliminary studies reported thus far, a revised form of the College Characteristics Index was produced in which 58% of the original items were retained, 13% were slightly modified, and 29% of the items were new. The revised form, at the time of writing this article, has been administered to approximately 1200 students distributed among more than 30 institutions. Further research and more intensive analyses have been planned. Before commenting on specific plans, however, certain broad values and implications in this psychological approach to the measurement of college environments are suggested.

One potential value, for example, is in institutional self-analysis. Administrators and faculty members should be able to learn something useful about the dynamics of the college environment from studying students' responses to the College Characteristics Index. Institutional press should have some clear relationship to institutional purpose. The objectives of a college are formal or explicit statements of intent: they indicate the directions in which a college means to influence the behavior of students. They find expression in curricula, practices, services, policies, and other aspects of the college environment. The press, as measured by the method described, constitute what Stern, Stein and Bloom (**9**) have referred to as an operational definition of objectives, or the implicit influence of the environment upon the students. Implicit press and explicit objectives should reinforce one another, for an institution should operate in reality the way it means to operate in theory. Consequently, a serious lack of congruence between implicit press and explicit objectives would suggest to faculty members and administrators that certain aspects of the environment ought to be changed in order to make the total impact of the institution more consistent or more effective. Pace (**6**) has commented elsewhere on the disintegrative effect of discrepancy between stated objectives and actual practices.

Some aspects of an environment can be changed more readily than others. The College Characteristics Index provides some direct indications of the psychological implications of various policies and practices. Roughly, one fourth to one third of the items in the Index state specific practices which an administration or faculty could more or less easily change if they did not like the implications. For example, being able to drop a course in which one is having difficulty, or to substitute another course for one which has been failed, is associated with Counteraction; insisting that students' reports or papers be neat, or giving students an assigned seat in class, or taking attendance regularly, is associated with Order. As the relationships among press variables and between these variables and institutional objectives as well as personal needs are established, the significance of such specific practices will become clarified. Other items in the Index are more indirect in their implications about the effect of various policies or practices. But the clues can be investigated and can thus be the starting point for serious discussions about the impact of the environment on the student and the relation of this impact to the intended objectives.

Another set of implications from this approach to describing college environments relates to the problem of assessment and prediction. Assessment studies have often failed because the situations or environments in which assessed modes of behavior were supposed to occur have been inadequately described or differentiated. The interaction between person and environment was not successfully predicted because the environment was not measured as analytically and systematically as the person. The whole field of college prediction studies provides a good example. The criterion is typically a grade point average. No fundamental improvement in predicting against this criterion has been made in the last 25 years. Prediction studies should be concerned with performance in the environment as a whole. The complexity of relationship between person and environment is inevitably obscured by the simplified and often inappropriate symbolism of correlation between scholastic aptitude test and grade point average. The press of a college environment represents what must be faced and dealt with by the student. It is possible that the total pattern of congruence between personal needs and environmental press will be more predictive of achievement, growth, and change than any single aspect of either the person or the environment.

It will be a long time before admissions officers or guidance counselors can benefit from the results of these more complex analyses. This requires the establishment of known relationships between kinds of persons and kinds of environments. But conceivably, advisers within an institution may be better able to help students find an effective and rewarding role within the operative environment of the college, or to see more clearly the ways in which environments need to be modified if different kinds of students are to grow within them most effectively.

As further steps toward refining and exploring the potential of this needs-press concept, as exemplified by the Activities Index and the College Characteristics Index, the following studies, among others, are in progress:

1. Statistical studies of the instrument (CCI), including test-retest reliabilities, correlation matrix of all scale scores, factor analysis, item covariances.

2. Comparative studies of types of items, subjective or impressionistic versus relatively objective or factual.

3. Studies of the relations between students' needs scores and the corresponding perception of press in the environment.

4. Analysis of perception of press among various sub-cultures within a complex institution.

5. Significance of congruence between Need and Press in determining successful performance and/or satisfaction in the college environment.

Summary

The present article has suggested that a college environment may be viewed as a system of pressures, practices, and policies intended to influence the development of students toward the attainment of important goals of higher education. The first draft of an instrument for measuring these influences systematically has been constructed. Data analyzed thus far reveal significant differences in the press of different college environments. The instrument itself, which has subsequently been revised, appears to be promisingly reliable and valid. In the long run, research which this type of instrument makes possible should increase understanding of the ways in which institutions make their impact upon students and provide a broader conceptualization for evaluating the effectiveness of higher education.

REFERENCES

1. Ebel, R. L. Procedures for the analysis of classroom tests. *Educ. psychol. Measmt.*, 1954, **14,** 277–282.
2. Edwards, A. L. *Edwards Personal Preference Schedule*. New York: The Psychological Corp., 1954.
3. Gardner, E. F., & Thompson, G. G. *Social relations and morale in small groups*. New York: Appleton-Century-Crofts, 1956.
4. McClelland, D. C., Atkinson, J. W., Clark, R. A., & Lowell, E. L. *The achievement motive*. New York: Appleton-Century-Crofts, 1953.
5. Murray, H. A. *Explorations in personality*. New York: Oxford Univer. Press, 1938.
6. Pace, C. R. Educational objectives. The integration of educational experiences. *Yearb. nat. Soc. Stud. Educ.*, 1958, **57,** Part III. Pp. 69–83.
7. Pace, C. R., & Stern, G. G. *A criterion study of college environment*. Syracuse: Syracuse Univer. Res. Inst., Psychol. Res. Center, 1958.
8. Stern, G. G. *Preliminary manual: Activities Index; College Characteristics Index*. Syracuse: Syracuse Univer. Res. Inst., Psychological Res. Center, 1958.
9. Stern, G. G., Stein, M. I., & Bloom, B. S. *Methods in personality assessment*. Glencoe, Ill.: Free Press, 1956.

Received June 5, 1958.

Journal of Educational Psychology
Vol.. 49, No. 5, 1958

PRESCHOOL IQs AFTER TWENTY-FIVE YEARS[1]

KATHERINE P. BRADWAY
Stanford University

CLARE W. THOMPSON[2]
State College of Washington

AND RICHARD B. CRAVENS[2]
Department of Health, Territory of Hawaii

In 1931, 212 children in the San Francisco Bay Area between the ages of 2 and 5½ years were given tests which later constituted Forms L and M of the 1937 Revision of the Stanford-Binet. Careful methods were developed and adhered to in the selection of these children to avoid biased sampling and to insure a representative group, because they were the California sample of the nationwide standardization of this revision (**9,** pp. 12–15, 18; **7,** pp. 6–7, 36–37).

Ten years later, in 1941, 138 of these children still in the area were administered Form L of the Stanford-Binet.[3] Thirteen of the original group were not included because as two-year-olds they had missed so many items at the lowest test level that the obtained mental age was a maximum expression of their intelligence, the true estimate not being determinable. Another 61 could not be located.

In 1956, 111 of the 1941 group were located and given Form L of the Stanford-Binet and the Wechsler Adult Intelligence Scale. Nine *S*s, although located, refused to participate; one *S* was in a mental hospital and not accessible for testing; another *S* was located in a community where no provisions could be made for testing; the remaining *S*s could not be located. A detailed report of the results of this follow-up and their theoretical implications is in the process of preparation. This is being interrupted now to present a summary of the actual IQ changes over the 25-year period because of the pending appearance of a new revision of the Stanford-Binet[4] and the appropriateness of having these retest data on the standardization group available at the same time.

Subjects and Procedure

The *S*s in the present study were administered Forms L and M of the Revised Stanford-Binet Scale 25 years previously, in 1931, in connection with its standardization. They were also examined with Form L of this same scale in 1941 (**3, 4, 5**). No interim contacts had been made. A comparison of the 1941 retest group of 138 *S*s with the total standardization group at these ages showed some upward selection, as the initial mean composite IQ of the retest group was 109.2 compared with 105.4 for the standardization group. Most of this difference was concentrated in the younger age groups,

[1] This investigation was supported by Research Grant M-1273 from the National Institutes of Health, Public Health Service.

[2] At Stanford University when participating in this research.

[3] The authors wish to express their debt of gratitude to Maud A. Merrill and the late Lewis M. Terman for having made their standardization data available for follow-up. They also wish to express their appreciation for the continued interest of Quinn McNemar, who has served as consultant for both follow-ups.

[4] Personal communication from Maud A. Merrill.

where the children for whom the test was too difficult had been eliminated. The distributions of paternal occupational level for the retest and standardization groups were similar at all ages.

The loss of *S*s between the 1941 and 1956 testings resulted in further selection with respect to initial mean IQ, which rose to 112.8 for the 1956 group with the elimination of 27 *S*s, 25 of whom were not located or refused to participate. Apparently availability for follow-ups was positively related to intelligence, although again the greatest discrepancy was for the group which had been two years of age at the initial testing.

In the current testing of 1956, Form L of the Revised Stanford-Binet and the Wechsler Adult Intelligence Scale were administered. (In one case the WAIS was incomplete. In another the 1931 IQ is not valid; this case was included only in comparisons with the 1941 results.) Of the 111 *S*s, 98 were tested by Cravens; two were tested by Bradway, who had administered all tests in the 1941 follow-up; the remaining 11 *S*s were tested by psychologists at colleges and universities located near the *S*s' current place of residence.[5] Mobility was less than had been anticipated: of the 122 *S*s located, only 24 lived more than 40 miles from where they had lived in 1931.

The retest group was comprised of 52 men and 59 women. This is the same proportion as in the 1941 testing. Inasmuch as no sex differences in test results were found in the standardization (**9**, p. 34), the results in the present study are given for the total group, combining both sexes.

TABLE 1

Obtained Stanford-Binet IQs by Year of Testing

	Mean IQ	SD
1931[a]	112.8	15.9
1941[b]	112.3	16.4
1956[b]	123.6	15.0

[a] Composite Form L and Form M.
[b] Form L.

Results

The degree of relationship between the initial composite IQ (Forms L and M) of these *S*s when they were in the age range from 2.0 to 5.5 (mean age 4.0) and Form L IQ when they were in the age range from 26.5 to 32.2 (mean age 29.5) is expressed by a Pearsonian *r* of .59. This compares favorably with the *r* of .65 found for the same group in the first follow-up after only 10 years. The correlation between the 1941 and the 1956 testings is .85.

All but 19 of the 110 *S*s showed a higher IQ on the 1956 testing than they did in 1931. This was not limited to any one segment of the IQ range. The mean increase was 10.8 points. The results in Table 1 show that this increase occurred between 1941 and 1956, that is between the early adolescent years and adulthood, since the mean IQs in 1931 and 1941 were approximately the same. The mean IQ increase from 1941 to 1956 was 11.3 points. The increases from 1931 to 1956 and from 1941 to 1956 are, of course, highly significant statistically with CRs of 8.1 and 13.7 respectively. The 1956 WAIS mean IQ of 108.9 ± 11.0 more nearly approximates the 1931 and 1941 Stanford-Binet mean IQs than does the 1956 Stanford-Binet.[6]

[5] The authors wish to acknowledge their indebtedness to the following psychologists who participated in this way: Steven K. Abe, Walter H. Brackin, Maurice Deigh, Robert Gray, George S. Ingebo, W. B. Lemmon, Jeanne Reitzell, Gordon and Virginia Riley, Alwyn Sessions, Paul S. Spitzer, and Helene R. Veltfort.

[6] The correlational values for the 1956 WAIS for this group are similar to those for the 1956 Stanford-Binet. Pearsonian *r* with 1931 Stanford-Binet is .64; with 1941 it is .80. The correlation between the two 1956 tests is .83.

These data are interpreted as invalidating the assumption that intelligence stops increasing at 16 years, on which was based the standard computation of an adult Stanford-Binet IQ. This is consistent with the findings of other recent investigators. Bayley (**1**) in the Berkeley Growth Study found increases in Stanford-Binet scores of the same *S*s up to the age of 17 years (the most recent administration), and on the Wechsler-Bellevue up to the age of 21 years (most recent administration). Moreover, a few 25-year scores available at the time of writing indicated that the ceiling of mental growth had not yet been reached by these *S*s. The interpretation of these results, of course, must take account of the possibility of practice effects from frequent retesting. Owens (**8**) found an increase on Army Alpha scores by *S*s who took the test initially at 19 years and again at 50 years. Bayley and Oden (**2**) in the most recent examination of Terman's Gifted Children found increases in scores on the Concept Mastery Test for *S*s covering a total range from age 20 to age 50 years, retested after a 12-year interval.

It seems obvious that a correctional factor of some sort should be applied to the adult Stanford-Binet IQs. Decision on the nature of such a correction for purposes of analyzing data of the present study is being deferred until the pending publication of the new revision. It may be noted, however, that whereas the IQs of 42% of the present group changed more than 10 points from 1931 to 1941, the IQs of 60% changed more than 10 points from 1931 to 1956. If the IQ is to be meaningful as an index, the average gain should be zero as it was between 1931 and 1941. When the IQs are equalized by taking account of the mean increase in IQ and subtracting 11 points from the 1956 value, the frequency of changes greater than 10 points is reduced to 42%. Similarly, 22% of the present group changed more than 15 points from 1931 to 1941, whereas 41% changed more than 15 points from 1931 to 1956. Equalizing the IQs as described above, the latter is reduced to 22%. From 1941 to 1956 the equalizing of the values reduces those changing more than 10 points from 58% to 18% and those changing more than 15 points from 28% to 7%.

Discussion

Beyond the problem of determining the optimal method to use in computing adult indices of intelligence is the question of what implications these data have for the nature of the mental growth curve and the theoretical question of terminal age of intellectual growth. A choice of method to use in arriving at the most useful index of adult intelligence requires the consideration of many factors. Wechsler's use of IQs computed by setting a mean of 100 and a standard deviation of 15 for each age is one solution. The central problem, however, is that intelligence, like so much of human behavior, is multifaceted. The point of view from which the question is approached depends upon what aspect is to be explored. Are we interested in an index showing one's place relative to others of the same age group, an index of one's location on a scale of development and decline, an assessment of one's various kinds of mental processes? Inasmuch as different abilities reach maturity at different ages and begin to decline at different ages, the particular abilities examined become deciding factors. For example, Corsini and Fassett (**6**) found from the testing of 1072 adults on the Wechsler-Bellevue that general intelligence declines from early to late maturity only if visual and motor factors are tested, and increases if the items are dependent on continued learning. We have in progress further research directed at determining what factors are associated with fluctuation in measured intelligence throughout the life span and the course of development and decline of different kinds of intelligence.

Summary

Of the 212 preschool children living in the San Francisco Bay Area who composed part of the standardization group of the 1937 Revision of the Stanford-Binet, 111 were retested as adults 25 years later. The sample was somewhat biased, with a tendency for *S*s at the lower IQ levels to be more difficult to locate or to refuse to participate. Thus the initial mean IQ of those retested was 112.8 as compared with 105.4 for the total group.

The Pearsonian *r* between adult and preschool Stanford-Binet IQs over the 25-year period is .59. The mean IQ showed a rise of 10.8 points in 25 years. This rise occurred in the years after early adolescence, there being no increase in mean IQ between preschool and adolescent testings. This is interpreted as invalidating the assumption that intelligence stops increasing at 16 years. Further research is in progress.

References

1. Bayley, Nancy. On the growth of intelligence. *Amer. Psychologist*, 1955, **10**, 805–818
2. Bayley, Nancy, & Oden, Melita H. The maintenance of intellectual ability in gifted adults. *J. Geront.*, 1955, **10**, 91–107
3. Bradway, Katherine P. IQ constancy on the Revised Stanford-Binet from the preschool to the junior high school level. *J. genet. Psychol.*, 1944, **65**, 197–217
4. Bradway, Katherine P. An experimental study of factors associated with Stanford-Binet IQ changes from the preschool to the junior high school. *J. genet. Psychol.*, 1945, **66**, 107–128
5. Bradway, Katherine P. Predictive value of Stanford-Binet preschool items. *J. educ. Psychol.*, 1945, **36**, 1–16
6. Corsini, R. J. & Fassett, K. K. Intelligence and aging. *J. genet. Psychol.*, 1953, **83**, 249–264
7. McNemar, Q. *The Revision of the Stanford-Binet Scale*. Boston: Houghton Mifflin, 1942
8. Owens, W. A. Age and mental abilities. *Genet Psychol. Monogr.*, 1953, **48**, 3–54
9. Terman, L. M., & Merrill, Maud A. *Measuring Intelligence*. Boston: Houghton Mifflin, 1937.

Received June 18, 1958.

Journal of Educational Psychology
Vol., 49, No. 5, 1958

THE ADEQUACY OF "MEANING" AS AN EXPLANATION FOR THE SUPERIORITY OF LEARNING BY INDEPENDENT DISCOVERY[1]

BERT Y. KERSH

University of Oregon

The hypothesis that learning through independent discovery is superior to learning by rote is well supported by existing research evidence (**5, 6, 7**). More recently, however, evidence has been published which suggests that attempts to direct the learner in the discovery process may also be successful without loss in retention or transfer (**2, 3**).

One reasonable explanation for the superiority of both the independent discovery and the directed discovery processes is that learning under either condition is more meaningful than in the case where the learner simply memorizes answers. Meaning is used in the following report in the cognitive sense of understanding or organization. More precisely, through the discovery process, in which the learner is forced to rely on his own cognitive capacities, he becomes cognizant of the relationships of the learning task to his previous experience, or to the pattern of relationships among the elements of the task. The superiority of such meaningful learning over rote learning is also well supported by research evidence (**1, 4**).

If meaningful learning is the key concept, it should make no difference whether learning occurs with or without direction, so long as the learner becomes cognizant of the essential relationships. However, it is very likely that some procedures of learning may be superior to others simply because they are more likely to cause the learner to become cognizant of the relationships.

The purpose of this research was to study the process of learning tasks involving arithmetical and geometrical relationships in order to determine whether or not the superiority of the discovery and directed-discovery procedure is adequately explained in terms of "meaningful learning," and, if not, to discover a more adequate explanation.

Description of the Tasks

Each *S* had the task of learning the following two rules of addition.

1. *The Odd-numbers rule.* The sum of any series of consecutive odd numbers beginning with one is equal to the square of the number of figures in the series. (For example, 1, 3, 5, 7, is such a series; there are four numbers, so 4 times 4 is 16, the sum.)

2. *The Constant-difference rule.* The sum of any series of numbers in which the difference between the numbers is constant is equal to one-half the product of the number of figures and the sum of the first and last numbers. (For example, 2, 3, 4, 5, is such a series; 2 and 5 are 7; there are four figures, so 4 times 7 is 28; half of 28 is 14 which is the sum.)

These rules can be learned by simply memorizing the task procedure as above. On the other hand, the learner can become cognizant of certain relationships to geometrical and arithmetical concepts which the two rules involve, in which case his learning will be more meaningful.

Consider first the Odd-numbers rule. In the first place, it is possible to relate the arithmetical concept of "squaring a number" to the geometrical concept of

[1] This study was supported by research grants from The Graduate School of the University of Oregon, Eugene, Oregon. Research assistance was ably provided by Harold K. Kozuma and Arthur O. Motz. An abbreviated version of this report was read at the joint meeting of the AERA and CERA in San Francisco, March 1958, under a different title.

"square." For example, "4^2" may be conceptualized as the arithmetical counterpart to the geometrical concept of a square having 4 units to the side.

Now, it can readily be shown that any series of odd numbers beginning with one can be rearranged in the form of a geometric square, as is illustrated in Fig. 1. In Fig. 1, each row of X's represents a number—the first row is "1," the second row is "3," and so on. The X's which are left outside the square can be fitted into the spaces indicated by the dashes inside the box.

The rule can also be related to the concept of the arithmetical average. The mean of any such series is the middle number; and the mean is always equal to the number of figures in the series. So the *mean times the number* is equivalent to the *number squared*. This relationship is suggested in Fig. 2.

Consider now the Constant-difference rule. The arithmetical and geometrical relationships involved in this rule are suggested in Fig. 3 and 4. Each figure essentially represents the original series inverted and added across; and it becomes apparent that the sum of the first and the last number, the sum of the second and the second-from-the-last number, and so on, is the same. The sum of the original series is, of course, one half the sum of the column of sevens formed in Fig. 3, and half the area of the rectangle thus formed in Fig. 4.

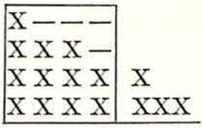

FIG. 1.

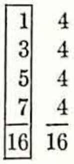

FIG. 2.

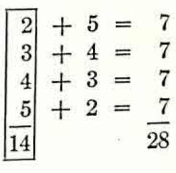

FIG. 3.

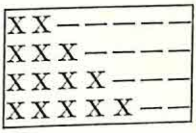

FIG. 4.

HYPOTHESES

The experiment was designed to test the following hypotheses:

1. It makes no difference in terms of retention or transfer effects whether the learner discovers the relationships which are essential to the understanding of a cognitive task independently or with external direction.

2. It is more probable that a learner becomes cognizant of the relationships which are essential to understanding a cognitive task when his attention is directed to the relationships than when his attention is directed to the task procedure alone, or when he is required to learn independently.

3. It follows from the second hypothesis that it is more probable that what is learned is remembered longer and transferred more effectively when the learner's attention is directed to the essential relationships than when his attention is directed to the task procedure alone, or when he is required to learn independently.

4. In the learning of tasks which involve both arithmetical and geometrical relationships, the learner most probably becomes cognizant of the former when conventional Hindu-Arabic symbols are used in directing his learning, and of the latter when more nearly iconic symbols are used.

PROCEDURE

One group of *S*s, called the "no-help"

group, was required to discover the rules without help. A second group, called the "direct-reference" group, was given some direction in the form of perceptual aids, such as those illustrated in Fig. 1–4 above, with accompanying verbal instructions which directed their attention to the perceptual aids. A third group, called the "rule-given" group, was told the rules directly and was given practice in applying them, without any reference to the arithmetical or geometrical relationships. The three procedures were called, as a group, the "teaching treatments."

In addition, there were two treatments called the "number treatments," which consisted simply of presenting the problems in either the more nearly iconic form, called the "X-form," or the conventional Hindu-Arabic form, called the "A-form."

The three teaching treatments and the two number treatments were combined to form six treatment combinations in a 2 × 3 factorial design. The treatment combinations were identified as follows: No help, A-form (No-help A); No help, X-form (No-help X); Direct reference, A-form (Dir-ref A); Direct reference, X-form (Dir-ref X); Rule given, A-form (Rul-giv A); Rule given, X-form (Rul-giv X).

A total of 60 college student volunteers from *E*'s two sections of Educational Psychology, taught in the Spring of 1957, formed the original sample. The *S*s were divided into six equal groups through the use of a table of random numbers.

The groups were compared in age, sex, grade level, and scholastic aptitude, and the differences were judged to be insignificant. However, nine of the original group were eliminated either because the recordings were defective or because they failed to complete the experiment. This left a total of eight in each group except one which had 10 remaining. Two more *S*s were eliminated from the group with 10 *S*s by the use of a table of random numbers so as to make the groups equal in number. The remaining 48 *S*s provided the data which were used in the analysis.

The procedure was to have each of the *S*s attempt to learn the two rules to the point where he could verbalize the rules and apply them in the solution of three different problems in succession. Tested individually, the *S*s were asked to "think aloud" during the learning period, and voice recordings were made.

Immediately following the learning period, a test consisting of 20 problems was given. Then, approximately four weeks later, the *S*s were asked to solve two simple problems and to fill out a questionnaire on their process of thinking.

The learning problems and the test problems were reproduced on separate pieces of paper or cardboard so that they could be presented one at a time. The experimental procedure may best be described step-wise, as follows:

Step 1. General instructions. The following instructions were given to each *S*.

> I am going to show you a series of addition problems one at a time, and you are to try to discover how to find the sum of each problem without adding in the usual manner.

Or, in the event the rule was to be given, the initial statement was as follows.

> I am going to tell you how to find the sum of two kinds of number series without adding, and then I will give you some practice in using the rules.

The remaining general instructions were the same for each experimental group.

> There are two types of problems, and there is a different rule for each of the two types.

At this point two sample problems were presented, as illustrated in Fig. 5.

> The first type, called the Odd-numbers problem, consists of a series of odd numbers

> beginning with one. Here is an example. (Point to the sample problem in the A-form.) Below the series you are given the number of figures in the series. In this case, there are five figures, so the "number" is five.
>
> Now look at the problem on the right. This problem is the same as the one on the left except that the numbers are represented in rows of X's. Each row represents a number. Do you see? (Explained, if necessary.)
>
> The second type problem is called the Constant-difference problem. This type consists of numbers which increase by a certain amount. The series may begin with any number and be of any length, but the numbers will always increase by a certain amount. In the example, the numbers increase by one. (Point to the sample problem in the A-form.) Again, on the right is the same problem in the X-form.

From this point the instructions were somewhat different for each treatment combination. For the two *no-help* groups the instructions were as follows.

> See if you can discover the rule for the Odd-numbers type problem first. Keep in mind that you are trying to discover a way of finding the sum of this type problem without adding in the usual manner. The problems I show you will all be examples of this type problem in the X-form (or conventional form).

At this point the first "learning" problem was presented. The learning problems appeared the same as the sample problem in Fig. 5, but were all in the appropriate number form.

> I want to record what you are thinking, so "think aloud" as much as possible. Tell me what you are thinking, in other words.

After *S* learned the rule to the Odd-numbers type problem, the Constant-difference problem was introduced by simply saying, "Now let's see if you can discover the rule to the Constant-difference type problem."

The specific instructions to the two *direct-reference* groups were as follows:

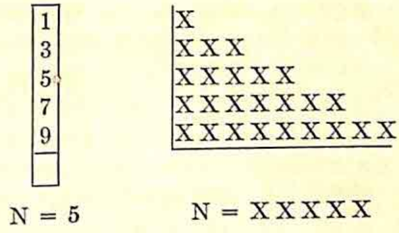

FIG. 5.

> See if you can discover the rule for the Odd-numbers type problem first. Keep in mind that you are trying to discover a way of finding the sum of this type problem without adding in the usual manner. The problems I show you will all be examples of this type problem in the X-form (or conventional form).
>
> In order to help you discover the rules, these examples provide you with additional hints. I'll show you these one at a time and explain the hints to you.

The first problem was presented at this point. Examples of the learning problems used with the Dir-ref X and Dir-ref A groups are illustrated in Fig. 1 and 2, respectively.

At this point, the instructions to the Dir-ref A group continued as follows.

> The problem will always be inside the box, and the hint will be to the right of the box where you see the column of fours.
>
> Notice that there are four 4's in the column, the same number of figures as there are in the problem series. Also notice that when you add the two columns, you get the same sum. Does that give you any ideas?
>
> I want a record of what you are thinking, so "think aloud" as much as possible. Tell me what you are thinking, in other words.

After *S* learned the rule to the Odd-numbers type problem, the Constant-difference problem was introduced in the following manner:

> Now let's see if you can discover the rule for the Constant-difference type problem. Again, the problem will always be inside the box, and the hint will be to the right of the box.

Essentially what we have done here is to take the problem series, invert it, and add across in the manner indicated. When we do this, you notice that we always get the same number, in this case 7. In other words, when you add the first number to the last number, you get 7; when you add the second number to the next-to-the-last, you get 7, and so on.

Also notice that the sum of the first column to the right is the same as that of the problem series, and the sum of the second column, the column of 7's, is twice the sum of the problem series. Now, does that suggest to you any ideas?

Starting again at the point at which the first problem is presented, the instructions to the Dir-ref X group continued as follows.

In these examples, the lines enclosing some of the X's and the dashes in a box are the hints. Notice that the number of dashes in the box is the same as the number of X's remaining outside. In other words, the total number of X's in the problem can be rearranged in the form of the box. Does this give you any ideas? Tell me what you are thinking. "Think aloud," in other words.

After the Odd-numbers rule was discovered, the Constant-difference problem was introduced as follows.

Now let's see if you can discover the rule for the Constant-difference type problem. Again the lines and dashes are your hints. Notice in this case that the number of dashes equals the number of X's. Essentially what we have done here is to invert the problem series and attach it to the original in the manner indicated to form the box. In other words, when you add the first row to the last you get 7; when you add the second row to the next-to-the-last, you get 7, and so on. In the end you have twice as many spaces within the box as there are X's in the original problem. Does that give you any ideas?

For the *rule-given* groups, the instructions were as follows:

The problems I will show you first are all examples of the Odd-numbers type problem in the X-form (or conventional form). The rule is to square the number of figures in the series to get the sum. The number of figures is 4. Four squared, or 4 times 4, is 16, which is the sum.

Now I will show you some more examples of this type problem. See if you can apply the rule correctly to each. Tell me what you are thinking.

After learning the Odd-numbers rule to the established criterion of three successful applications in succession, the Constant-difference type problem was introduced as follows.

Now let's see if you can learn the rule to the Constant-difference type problem. The rule is to add the first and the last numbers in the series, multiply by the number of figures in the series, and then divide by two. With the first example, add the first and the last numbers (2 and 5 are 7), multiply by the number (4 times 7 is 28), and divide by two (28 divided by 2 is 14). Fourteen is the correct sum. Now you practice the rule on these other examples.

Step 2. The learning period. The learning period was considered to begin immediately after the instructions in Step 1 were given, and to end immediately after the *S* successfully applied an acceptable rule to three problems in succession. An acceptable rule was defined as one which could be used in the solution of the problems as presented in the learning period, even though it might be inadequate for use with the problems as presented in the test period. For example, a rule which was frequently considered is to multiply the middle number in the series by the number of figures. This was adequate during the learning period because the problems included all the numbers; however, the test problems included only the first three numbers and the last of the series, which made the rule awkward to apply. This criterion forced *E* to deviate slightly from the experimental procedure in some cases, as will be described below.

Six different problems were used in the learning period, and they were basically the same for each group. The general procedure was to wait until S had developed a possible solution with one problem, then to give him the next problem, telling him to try it on another example. This procedure was continued until S was successful. S was permitted to manipulate the cards himself if he wished, to back-track, or to look at several problems simultaneously. Also, E repeated the instructions when asked, and answered questions regarding the task. Otherwise, E limited his remarks to those which were intended to give support and encouragement, such as, "Try your idea on the next problem," "You are doing very well," and "See if it works." In order to keep S thinking aloud, E would ask, "What are you thinking now?" whenever S was silent for more than 10 or 15 seconds.

In the event that S discovered a rule which appeared to work satisfactorily but which was not the intended rule, E first attempted to encourage him to continue searching by saying, "That's fine, now see if you can discover another rule." In the event that S perseverated, E would shift to the problems used during the first test period, and, as soon as S became aware of the inadequacy of his rule, would shift back again to the learning problems.

There were some individuals, particularly in the no-help groups, who were simply unable to discover the intended rules within the time period of 60 to 90 minutes scheduled for each S. In this event, the learning period was terminated.

Step 3. The first test period. Immediately following the learning period, S was given 20 problems to solve, five each of the following: Odd-numbers problems (A-form); Odd-numbers problems (X-form); Constant-difference problems (A-form); Constant-difference problems (X-form).

The 10 problems of each type were actually five problems presented once in the A-form and once in the X-form. The problems were grouped by type and numerical form and were clearly identified as such. The procedure was to present the Odd-numbers problems first, then the Constant-difference problems. Furthermore, the Ss who learned the rules using problems in the A-form were presented the test problems in the A-form first.

Step 4. The retest period. A retest was given to all Ss four to six weeks after the first test. The second test consisted of two problems which could most easily be solved by using the two rules which the Ss attempted to learn during the learning period, and a series of questions on their process of thinking. The problems and questions were completed under supervision, all in one day. The instructions were to record all scratchwork in spaces provided for this purpose, and to write the answers to the questions as clearly and completely as possible. The problems were as follows: (*a*) What is the sum of the first 35 odd numbers, and (*b*) what is the sum of *all* the first 35 numbers? The questions which followed each problem were the following:

> Did you add the first 35 (odd) numbers to get your answer? (Yes or no) If your answer is no, explain how you obtained your answer.
>
> Did you try to recall the rule you learned (or attempted to learn) under our direction several weeks ago? (Yes or no)
>
> Were you successful in recalling the rule? (Yes or no)
>
> Describe how you recalled or attempted to recall the rule.

Analysis of the Results

Not all Ss succeeded in learning the two rules as stated above to the established criterion. Some of those who failed did manage to discover other workable variations of the rules which were judged to be acceptable. Others, however, failed to learn any workable rule for one of the two tasks before the practical limitations of time forced E to terminate the learning period.

TABLE 1

NUMBER WHO LEARNED EACH RULE (OR AN ACCEPTABLE VARIATION) TO CRITERION DURING THE LEARNING PERIOD

Rule[a]	No-help		Dir-ref		Rul-giv		All treatments
	A	X	A	X	A	X	
	Odd-numbers rule						
1. n^2	7	6	8	7	8	8	44
2. m^2	1	0	0	0	0	0	1
3. mn	0	1	0	0	0	0	1
4. $[\frac{1}{2}(a + l)]^2$	0	0	0	1	0	0	1
5. none	0	1	0	0	0	0	1
All rules	8	8	8	8	8	8	48
	Constant-difference rule						
1. $\frac{1}{2}n(a + l)$	1	0	7	7	8	8	31
2. mn	1	1	0	0	0	0	2
3. none	6	7	1	1	0	0	15
All rules	8	8	8	8	8	8	48

[a] In this column, n = number of figures in series; m = mean, or median of series; a = first number, and l = last number in series.

Table 1 shows the number of those who learned the rules and each of the acceptable variations, and those who failed to learn any workable rule in each case.

Those who failed to learn an acceptable rule were retained in the experiment and retested along with the others four weeks later, in anticipation of the possibility that their performance on the retest problems and their responses to the retest questionnaire would reveal useful information for further research. This indeed proved to be the case.

The purpose of the 20-item test given immediately after the learning period was to detect differences among the *S*s in their achievement, if any. All *S*s who learned the stated rules to criterion scored perfectly on the test, and those who failed to learn the acceptable rules were, of course, unable to solve the problems without adding.

On the retest given four weeks later, the datum of primary interest was the method used in solving the problems, not whether or not the correct answer was provided. Table 2 shows the methods used on the retest and the number of *S*s in each group who used each method.

A comparison of Table 2 with Table 1 reveals the changes in the learned methods and the methods used on the retest, but does not indicate whether or not the methods that were used on the retest were the same as those which were developed during the learning period. It is apparent that there were at least three possibilities: (*a*) the same method was used, (*b*) a different method was used, or (*c*) the problem was solved by simple addition, which would indicate complete failure to recall or transfer the rule if learned. Table 3 shows the number of *S*s in each of the teaching treatment groups who, on the retest, used the same method, some other method than that which they had learned, or simple addition. The *S*s who failed to learn any acceptable rule in the learning period were placed in the "added" category if they added on the retest, and in the "other" category if they attempted some other procedure on the retest.

The results presented above fail to support the hypothesis that the *Ss* whose attention is directed to the relationships which are essential to understanding remember their learning longer and transfer it more effectively than do the *Ss* of the other two treatment groups (Hypothesis 3, above). The prediction was that the direct-reference group would be superior. Instead, although the obtained differences are not highly reliable, they are consistent with previously published data to the extent that they suggest that the independent discovery procedure is superior to learning by rote.

The reader's attention is directed to the fact that although 13 *S*s in the no-help groups failed to learn an acceptable rule for the Constant-difference problem during the learning period, there were only four who added on the retest, and 10 who

TABLE 2

METHODS USED ON THE RETEST

Method[a]	No-help		Dir-ref		Rul-giv		All number treatments		All teaching treatments		
	A	X	A	X	A	X	A	X	No-help	Dir-ref	Rul-giv
	Problem 1. Sum of the first 35 odd numbers										
1. n^2	6	6	2	2	5	1	13	9	12	4	6
2. $\frac{1}{2}n(a + l)$	0	1	0	1	0	0	0	2	1	1	0
3. $[\frac{1}{2}(a + l)]^2$	0	0	1	1	1	0	2	1	0	2	1
4. mn	2	0	0	0	0	0	2	0 [b]	2	0	0 [b]
5. Not acceptable	0	0	3	1	1	1	4	2	0	4	2
6. By addition	0	1	2	3	1	6	3	10	1	5	7
Chi square	Not applicable						1.39*		11.18**		
	Problem 2. Sum of all the first 35 numbers										
1. $\frac{1}{2}n(a + l)$	3	2	3	4	4	0	10	6	5	7	4
2. mn	3	2	0	0	0	0	3	2 [b]	5	0	0 [b]
3. Not acceptable	1	1	3	2	1	0	5	3	2	5	1
4. By addition	1	3	2	2	3	8	6	13	4	4	11
Chi square	Not applicable						1.36*		3.86*		

[a] In this column, n = number of figures in series; m = mean, or median of series; a = first number, and l = last number in series.

[b] Frequencies above and below line combined for chi-square analysis.

* Not significant at the 0.05 level.

** Significant at the 0.05 level.

TABLE 3

CORRESPONDENCE OF METHODS USED ON THE RETEST WITH THOSE LEARNED DURING THE LEARNING PERIOD

Method	No-help		Dir-ref		Rul-giv		All number treatments		All teaching treatments		
	A	X	A	X	A	X	A	X	No-help	Dir-ref	Rul-giv
	Problem 1. Sum of the first 35 odd numbers										
Same	6	5	2	3	5	1	13	9	11	5	6
Added	0	1	2	3	1	6	3	10	1	5	7
Others	2	2	4	2	2	1	8	5	4	6	3
Chi square	Not applicable						3.48*		8.21*		
	Problem 2. Sum of all the first 35 numbers										
Same	3	2	3	3	4	0	10	5	5	6	4
Added	1	3	2	2	3	8	6	13	4	4	11
Other	4	3	3	3	1	0	8	6	7	6	1
Chi square	Not applicable						3.03*		9.99**		

* Not significant at the 0.05 level.

** Significant at the 0.05 level.

TABLE 4

Number Who were Judged to be Cognizant of Each Type of Relationship During the Learning Period

Type Relationship	No-help		Dir-ref		Rul-giv		All Treatments
	A	X	A	X	A	X	
	Odd-numbers rule						
Arithmetical	1	1	1	0	0	0	3
Geometrical	0	2	0	4	0	1	7
Both	0	0	0	1	0	0	1
Neither	7	5	7	3	8	7	37
	Constant-difference rule						
Arithmetical	3	1	6	0	0	0	10
Geometrical	0	0	0	2	0	0	2
Both	0	0	0	0	0	0	0
Neither	5	7	2	6	8	8	36

used acceptable methods. In the case of the other two teaching treatment groups, the number who added increased, and the number who used acceptable rules decreased rather markedly. This finding suggests that as a result of their experience during the learning period, the *S*s in the no-help groups were motivated to continue learning afterwards and those treated otherwise were not.

The superiority of the discovery procedure may be better explained in terms of motivation, then, than in terms of understanding. This suggestion is given added support by the following data.

In an attempt to determine the number of *S*s in each experimental group who actually became cognizant of the arithmetical, the geometrical, or both types of relationship, the typed transcriptions of the voice recordings made during the learning period were carefully examined. Any *S* was judged to be cognizant of a relationship if he verbalized it at any point in the process of his learning. As a result of *E*'s continued efforts to stimulate each *S* to think aloud during the learning period, very complete records were obtained. In addition, each *S* who failed to volunteer an explanation of the rule was asked to do so at the end of the period. In most cases of the latter type, the *S*'s response was that he could not provide an explanation. Table 4 shows the results of this analysis.

Most striking is the fact that only one-quarter of the entire group of 48 *S*s was judged to be cognizant of one or both types of relationship to each of the two rules. The small number of *S*s precludes any statistical check of the reliability of the obtained difference in the experimental groups. However, accepting the obtained data as reliable, the results appear to support the prediction that the direct-reference treatment would produce the greatest incidence of understanding. The next highest frequency of understanding was in the no-help group, and even in the rule-given group there was at least one *S* who understood the odd-numbers rule. Furthermore, with only two exceptions, both in the no-help group, the type of relationship discovered corresponded to the type that was predicted from the number treatment. When the problems were presented in the A-form, the arithmetical relationship was discovered, and when the X-form was used, the geometrical relationship was discovered. The data, therefore, tend to support Hypotheses 2 and 4.

The data presented in Table 4 do not preclude the possibility, however, that the *S*s who were not cognizant of the relationships at the end of the learning period subsequently did become cognizant of them. The most adequate test of the efficacy of understanding in learning is to determine how those who actually verbalized the relationships performed on the retest problems. Table 5 shows the number of those who, on the retest, used the same rule they learned, those who simply added, and those who used some other method than that learned.

Again, the dearth of data does not per-

TABLE 5

METHODS USED ON RETEST OF THOSE WHO WERE JUDGED TO BE COGNIZANT OF THE ESSENTIAL RELATIONSHIPS

Methods	No-help		Dir-ref		Rul-giv		All Groups
	A	X	A	X	A	X	
	Problem 1						
Same Rule	1	3	0	2	0	0	6
Added	0	0	0	1	0	1	2
Others	0	0	1	2	0	0	3
All Methods	1	3	1	5	0	1	11
	Problem 2						
Same Rule	2	1	3	0	0	0	6
Added	1	0	0	2	0	0	3
Others	0	0	3	0	0	0	3
All Methods	3	1	6	2	0	0	12

mit a conclusive test of the first hypothesis. However, since only about one-half of those who learned acceptable rules with understanding during the learning period succeeded in retaining and transferring their learning four weeks later, the data do not offer very conclusive support of the meaning theory.

DISCUSSION

Perhaps the data which have been presented are sufficient to suggest the inadequacy of the "meaning" explanation, but they fail to communicate the impression of markedly increased motivation or interest in the task which characterized many of the *S*s in the no-help group as compared with many of those in the other groups. The difference in motivation was reflected in their written comments on the retest and verbal reports to *E*. For example, one *S* in the no-help group reported that he was so intrigued with his success in discovering the rules that he told his friends of his experience and tested their ability. Others in the no-help group who failed to discover the rule told of their efforts to learn the rule, even going so far as to look up the algebraic formula in the library. On the other hand, one *S* in the rule-given group complained that *E* had not instructed him to remember the rules, so he had promptly forgotten them. Others in the rule-given and direct-reference groups commented on their confusion during the learning period in explaining their inability to recall the rules. The evaluation of the *S*s' comments was admittedly highly subjective, but it may better suggest the nature of the motivation which is offered as the more adequate explanation for the superiority of the discovery procedure.

Of the various descriptive concepts of human motivation known to this present writer, Allport's concept of functional autonomy best describes the motivation that was developed by those in the no-help group. Whereas, at the beginning of the learning period, all the participants were assumed to be alike in that they were motivated primarily by such extrinsic factors as the approval of *E*, at the end of the learning period, the motivation of many of those in the no-help group appeared to be independent of the experimental or instructional situation. Presumably, the motivating power is of the type that lies in acquired interest or ego involvement in a task, and develops to the extent that the individual relies on his own cognitive capacities in learning.

The findings place the teacher in somewhat of a dilemma. He must decide which is the more important outcome of learning, (*a*) maximum understanding, or (*b*) maximum motivation to continue learning. However, further research on the problem may reveal ways of directing the discovery process without inhibiting the development of autonomous motivation. Teachers should continue to strive to guide the learning of their students but should refrain from giving answers directly.

Finally, the findings suggest that more consideration be given to the influence of the stimulus materials in directing the thought processes of the learner. Particularly recommended is a careful study of the educational "toys" that are used in some primary classrooms to develop number concepts, and the practice exercises in workbooks and on worksheets which are intended to direct the learner's attention to relationships.

Summary and Conclusions

An experiment was performed with college students in order to test the premise that learning by independent discovery is superior to learning with direction because the learning is more meaningful in the former case. "Meaningful learning" was used in the cognitive sense of understanding or organization. The Ss were required to learn two arithmetical tasks to a common criterion. A two-factorial design was used which involved three "teaching" treatments (no-help, direct-reference, and rule-given), and two "number" treatments (A-form and X-form). Approximately four weeks later a retest was administered which was designed to test the ability to recall and apply the learned rules to a somewhat different type of problem from that used in the learning period. The Ss' self-reports of their learning process and methods used on the retest were analyzed.

The conclusion is that the superiority of the discovery procedure of learning over procedures of learning with external direction is not adequately explained in terms of "meaningful learning." The results of this experiment suggest that when the learner is forced to rely on his own cognitive capacities, it is more likely that he will become motivated to continue the learning process or to continue practicing the task after the learning period. Consequently, the learning becomes more permanent and is more effectively transferred than when the learner is not so motivated.

References

1. Brownell, W. A., & Moser, H. E. Meaningful versus mechanical learning: A study in Grade III subtraction. *Duke Univer. Res. Stud. in Educ.*, 1949, No. 8.
2. Craig, R. C. *The transfer value of guided learning.* New York: Teachers Coll., Columbia Univer., 1953.
3. Craig, R. C. Directed versus independent discovery of established relations. *J. educ. Psychol.*, 1956, **47**, 223–234.
4. Katona, G. *Organizing and memorizing: Studies in the psychology of learning and teaching.* New York: Columbia Univer. Press, 1940.
5. McConnell, T. R. Discovery vs. authoritative identification in the learning of children. *Univer. Iowa Stud. in Educ.*, 1934, **9**, No. 5, 11–62.
6. Swenson, Esther., Anderson, G. L., & Stacey, C. L. *Learning theory in school situations.* Minneapolis: Univer. Minnesota Press, 1949.
7. Thiele, C. L. The contribution of generalization to the learning of the addition facts. *Teach. Coll. Contrib. Educ.*, 1939, No. 763.

Received July 11, 1958.

THE JOURNAL OF

EDUCATIONAL PSYCHOLOGY

Volume 49 December, 1958 Number 6

THE TRANSFER VALUE OF GIVEN AND INDIVIDUALLY DERIVED PRINCIPLES[1]

G. M. HASLERUD AND SHIRLEY MEYERS

University of New Hampshire

While there is general recognition of the value of organized learning for memory and transfer, differences on how that organization should be taught and attained lead to quite different theories of transfer. On the one hand are those who decry outside direction of learning when one is interested in transfer. Katona (**7**) using card and geometic puzzles found that his memorization group was significantly poorer on invention and transfer to new problems than his "Help" group that had only examples. He concluded "... that formulating the general principle in words is not indispensable for achieving application," (**7**, p. 89) but he was unwilling to say that learning of principles in words is always less efficient than by example. He put teaching the result as the worst method, teaching by stating the principle as intermediate, and teaching by example as best. However, Hendrix (**5**) found that with a mathematical principle those groups that discovered the principle independently and left it unverbalized exceeded those who discovered and then verbalized, and both exceeded in transfer those who had the principle stated for them and then illustrated.

Opposed to Katona and Hendrix are those like Craig who concluded: "The more guidance a learner receives, the more efficient his discovery will be; the more efficient discovery is, the more learning and transfer will occur" (**1**, p. 72). In a further study with college groups and with the same method of having the *S* pick out that alternative among five which does not fit a principle (**2**) he verified that significantly more such problems were solved when the principle was stated above it than when the *S* was given only the instruction that one of the five items did not belong. One should note, however, that he found no difference between his groups on transfer to new principles nor was there any difference in retention after 3 or 17 days, although at 31 days the difference favored the directed group.

While Craig's experimental results actually gave little or no support to his claim that guidance is desirable for transfer, more serious opposition to Hendrix and Katona came from Kittell (**8**). He found that "intermediate direction" (starting a principle) was significantly superior to both the "minimal direction" (told only that one of five alternatives would not fit) and "maximal direction" (*E* told the principle and worked out the answer for *S*), with minimal direction definitely the inferior method. His subjects were sixth graders while Craig's were college students. That difference in age and educational level may explain the contrast in results. Also Kittell's low number of successful solutions (means only 4.59 for intermediate direction and 1.93

[1] This research was partially supported by the Central University Research Fund of the Graduate School of the University of New Hampshire.

for minimal direction out of 15 principles) suggests that the problems based on linguistic arrangements and meanings may have been too difficult for the *S*s. If that were the case, then following directions in the stated principle was about the only way to solve the problem when unprovided with sufficient apperceptive mass and experience. Haslerud (**4**) found that while naive rats transferred anticipatively from forced turns near the goal into prior free units of a maze just as well as when those goal turnings had been established by trial and error, only active trial and error cul-de-sac elimination in the goal region could readjust an established pattern in the prior free units. If a similar limitation on effectiveness of guidance is present in human *S*s, then one might expect any advantage primarily in young *S*s and that mainly on their initial learning but none for memory and transfer where *S*s have sufficient background to derive a solution themselves.

While the Katona and Hendrix concept of how to get maximal transfer seems to have face validity, at least for adults, their controls and statistical supports are unsatisfactory. When one draws his conclusions on the basis of one principle, e.g., the sum of the first *n* numbers, a question remains of how much of the conclusion is a function of the particular problem used or the selection of individuals for the various groups. More convincing differentiation of principle given from principle derived would seem to require homogeneously varied problems posed in quantity to the same individuals. A likely material has been found in an extension of the familiar cryptogram "Come to London" in the Stanford-Binet. An unpublished pilot study by the junior author under the senior author's supervision indicated an advantage for memory of the independent solving of such coding principles. The present study extends a similar method to transfer. The hypothesis tested is that principles derived by the learner solely from concrete instances will be more readily used in a new situation than those given to him in the form of a statement of principle and an instance.

Procedure

Subjects for the experimental group were 76 members, ranging from freshmen to seniors, of two general psychology classes at the University of New Hampshire. The control group of 24 students in another psychology class ranged from sophomores to seniors.

The experimental groups were each given two coding tests, the second being administered one week after the first. The control group was given only the second test. All tests were administered by the senior author.

The first test composed of 20 coding problems was designed to give the students two types of experience: (*a*) problem solving with specific directions for deciphering the code printed above each problem, and (*b*) problem solving with no directions given. The first part of each problem was the four-word sentence "They need more time," followed by the same sentence in code. A different code was used in each problem. The second part of each problem was the four-word sentence "Give them five more," which the *S*s were asked to translate into the code for that problem. The given and derived problems were alternated so that the *S* would solve approximately equal numbers of each kind. As a control for differences between the codes, there were two test forms, A and B. The same codes were used in both, but those for which directions were given in form A had to be deciphered by the *S* in form B, and vice-versa. The problems were arranged in approximately the apparent order of difficulty. Examples of moderately easy coding rules are: "For each letter of the sentence write the letter that follows it in the alphabet." "Write the first two letters of each word and then the last two letters of each word." To introduce the test the senior author told the *S*s that the test was an experiment in cryptography. He wrote an illustrative code on the blackboard and purposely worked it out partly incorrectly to encourage remonstrances from the group that a system or principle was possible. *S*s were asked to solve the problems in the order they appeared on the test and to do as many as they could in the time allotted. Since the 45 minutes allotted was ample time for all but one or two students in each group, the test was essentially a power rather than a speed test. The *S*s were not told that they would be retested on the same material.

The second test printed only in one form was given to both the experimental and control groups. Again, the 20 codes used in the first test were used, but instead of the common sentence of Test 1, there were 20 different English sentences 14 to 18 letters in length followed by four translations into code. Only one translation was correct, and the *S*s were asked to check it. They were told that the other three were simply letters arranged in random order. They were not told that numbers had been assigned to letters of the alphabet and that letters for two of the four codes had been selected according to the order in which those numbers appeared in a list of random numbers. The third false code was composed of letters of the English sentence arranged according to random numbers. The order in which the four codes followed the sentence and the order in which the problems were arranged on the test were also random. No mention was made of the previous test, nor was the purpose of the test told until both tests had been given and the results compiled.

Results

The data for each individual in the experimental group consisted of four scores: (*a*) Number of correct codings on Test 1 problems where the rule was given, hereafter called G_1 scores. (*b*) Number of correct codings on Test 1 problems where the coding principle had to be derived by the *S*, hereafter called D_1 scores. (*c*) Correct alternatives for those codes in Test 2 that had been G type in Test 1. (*d*) Correct alternatives for those codes in Test 2 that had been D type in Test 1. In the control group the score was the total number of correct alternatives on Test 2. Any coding was considered correct if no more than 1 of the 16 letters was wrong, since carelessness rather than lack of understanding of the principle was probably responsible for the lone error.

Since there was no difference between their results, the two experimental classes were combined. The analysis of results, however, was carried through separately for Forms A and B of Test 1 because a difference significant at the .05 level indicated that the 10 odd and the 10 even problems had not been exactly equated for difficulty. Nevertheless, the direction of results for both A and B groups showed equally high differentiation of G and D situations.

Test 2 performance of the experimental group was significantly different from that of the control group. The means, 15.74 and 10.75 respectively, differ beyond the .001 level. Apparently something is transferred from the Test 1 experience.

The crucial comparisons are between the G and D kinds of problems. For both Forms A and B on Test 1, significantly more G problems were correctly coded: 8.86 and 8.36 against 5.86 and 4.88 for G and D respectively. The results for Test 2 a week later are given in Table 1. If the differences are added algebraically to the

TABLE 1

Difference in Number of Problems Successfully Coded Between the Transfer Test (Test 2) and the Initial Learning (Test 1) with Each Individual as His Own Control

	N	$\bar{x}$ diff	∂_{diff}	$\partial_{\bar{x}diff}$	t	signif.
D_2-D_1 Problems						
Test A	36	2.83	3.34	.56	5.06	$p < .001$
Test B	40	2.50	3.30	.52	4.81	$p < .001$
G_2-G_1 Problems						
Test A	36	−0.83	2.37	.38	2.17	$p < .05$
Test B	40	−1.03	2.21	.35	2.94	$p < .01$

Test 1 scores given in the previous sentence, one obtains the nearly equal transfer scores of Craig's experiment (**2**). But since each individual was his own control for both G and D problems on Tests 1 and 2, it is legitimate to use the subtraction method to find the standard error of the difference for paired observations. The correct identification of those codes which had been D type on Test 1 increased 46% while those which had been G decreased 10%. Both changes are significant, at the .001 and .05 to .01 levels respectively. There is reason to think that both curtailing time to make Test 2 a speed test and increasing time to greater than a week between the learning of the codes on Test 1 and the transfer on Test 2 would accentuate the differences.

Discussion

This experiment has added strong support to the contention of Katona and Hendrix that independently derived principles are more transferable than those where the principle is given to the student. Even though *S*s produced more correct codings on the original learning when the principle was stated for them, on the "pay-off," or "applying" to use Katona's term, the advantage definitely passed to those principles derived by the student himself. Fast and accurate learning or performance under immediate guidance is no guarantee of transfer to new problems without such support. From Craig's and our experiments the conclusions just stated are supported by results on college students, but testing of grammar level students by principles of a more suitable level of difficulty than used by Kittell (**8**) might show a wider application. Our coding method could be easily adapted for that purpose.

The obtained results of this experiment do not follow from inadequate controls. The alternate Forms A and B allowed each principle to be given (G) and derived (D). Individual differences with respect to problem solving in the *S*s were ruled out since each person responded to 10 G and 10 D problems on Test 1 and the follow-up of each of these on the transfer Test 2. The control group's much poorer performance on Test 2 indicated that a genuine transfer function was present. Making time on each test practically unlimited pushed the G and D types of presentation to their limit as power tests.

Two possible weaknesses in the transfer Test 2 need to be examined. With four alternatives for each problem, a chance score would average 5. The control group had 10.75 problems correct; this showed good adaptation to the test but significantly less than the 15.74 of the experimental groups. The question whether the better performance of the experimental group was just the result of a second session of practice on coding problems can probably be answered by reference to the study by Warren (**9**). He found that adults on letter-symbol substitution rapidly attain a plateau on transfer problems because of "learning sets" from early childhood. Coding is in that class of simple activities for adults where experience and practice as such make little difference after the first 10 minutes. Even if one took the maximum change of 37% during Warren's 16 five-minute periods, it would be less than the nearly 50% advantage of our experimental group over the control group. The second possible weakness arises from the randomized construction of the false alternatives of the transfer test. It is conceded that a person might try to solve the problems by excluding the three alternatives because of their random characteristics rather than by trying to recognize and verify some consistent principle in the one true alternative. However, the principles must have played a significant role in the solutions because without them the results of the control and experimental groups would have been equal since they had the same instructions and equal opportunity to use this abortive device.

The theories of transfer found in current educational psychologies are inadequate to explain the present experiment. The senior author plans to develop in another place a theory that transfer is fundamentally an anticipative rather than a perseverative function and that to get transfer one must always counteract the finality of a goal (**3**). A stated principle to some extent, and even more Kittell's "maximum guidance" of *E* doing the problems for *S* after giving him the principle, practically stops transfer, like other goals. Hendrix (**6**) states from Thorndike that only 5% of high school students have language ability sufficient to receive a ready-made sentence and find readily illustrations in their own background to provide the prerequisite to meaning. If the results of the present experiment can be verified for a wider range of ages and apperceptive masses, then the implications for a direct attempt to teach for transferable principles can not be neglected.

Summary

The educationally important question of how much guidance is desirable if one is interested in transfer was tested experimentally by a new use of coding. Each of 76 college students as his own control translated into 20 different codes a common four-word sentence, with the rule given for half of the problems and required to be derived solely from example for the other half. As in previous studies on initial learning, the *S*s did significantly better on those problems with the rule given. However, a week later on a multiple-choice transfer test consisting of 20 different sentences, one for each of the 20 coding principles of the first test, the selection of the adequate code from three specious ones made by randomizing letters gave very different results. The scores were significantly increased for those problems which had formerly been derived as contrasted with a significant decrease for those problems where the rule had formerly been given. A control group of 24 college students given only the second test proved by significantly poorer performance than the experimental group the value of transfer from the first test. The results give strong support to the postulate of Hendrix that independently derived principles are more transferable than those given. The apparent contradiction with Kittell's study of children was explained by the smaller apperceptive mass in the child, and the prediction was hazarded that as naivety is lost, the probability of transfer from learning which is minimally directed is increased.

References

1. Craig, R. C. *The transfer value of guided learning*. New York: Teachers College, Columbia Univer. Bureau of Publications, 1953.
2. Craig, R. C. Directed versus independent discovery of established relations. *J. educ. Psychol.*, 1956, **47**, 223–234.
3. Haslerud, G. M. Properties of bi-directional gradients at subgoals. *J. genet. Psychol.*, 1950, **29**, 67–76.
4. Haslerud, G. M. Anticipative transfer of mechanically guided turns. *J. exper. Psychol.*, 1953, **45**, 431–436.

5. Hendrix, Gertrude. A new clue to transfer of training. *Elem. Sch. J.*, 1947, **48,** 197–208.
6. Hendrix, Gertrude. Prerequisite to meaning. *Math. Teacher,* 1950, **43,** 334–339.
7. Katona, G. *Organizing and memorizing.* New York: Columbia Univer. Press, 1940.
8. Kittell, J. E. An experimental study of the effect of external direction during learning on transfer and retention of principles. *J. educ. Psychol.*, 1957, **48,** 391–405.
9. Warren, J. M. Intertask transfer in code substitution learning. *J. genet. Psychol.*, 1956, **89,** 65–70.

Received August 18, 1958.

Journal of Educational Psychology
Vol. 49, No. 6, 1958

OCCUPATIONAL LEVEL AND THE PRIMARY MENTAL ABILITIES[1]

K. WARNER SCHAIE
University of Nebraska

Thurstone's S.R.A. Primary Mental Abilities test (PMA) is frequently used as the intelligence test component of a battery given for guidance purposes. A reason for this use is the common inference that a study of the separate and presumably independent scores for the different abilities will yield clues to predict future success in certain school courses and vocations. Since intellectual functioning is generally found to be an important determinant in predicting successful performance it would be of interest to validate assertions that one could go a step further and predict differential success for a given type of vocational choice.

A scrutiny of the PMA literature suggests that validation studies have been concerned primarily with the correlation of the Primary Mental Abilities with a variety of achievement tests. Examples of such studies are reported in the test manual for the relation of the PMAs with the Stanford Achievement test (**5**) and with the Iowa Tests of Educational Development (**4**). Other work has related the PMAs to the United States Employment Service General Aptitude tests (**2**). These studies, done primarily with high school populations, conclude that the PMAs are fairly good predictors of current achievement and are useful for guidance purposes.

None of these studies provide any information, however, upon success in predicting actual occupational choice. The present inquiry attempts to fill this gap indirectly by examining a group of adult individuals who have made a firm occupational choice to see whether they could be differentiated in terms of their PMA scores as to the type of occupation selected.

Hypotheses

The PMA manual was consulted to see what kind of predictions were suggested by the test authors and others in relating performance on the PMAs to activities required in various occupations. One purpose of the PMA profile, for example, is its use for estimating the individual's general level of intelligence. Young people planning to go to college are presumed to require above average standing on most of the abilities, but particularly on Verbal-meaning (V) and Reasoning (R). People whose occupational choice results in professional types of activity would therefore be expected to show high performance on all abilities but should show particular elevation on V and R. Space ability (S) is presumed to be important for occupations like electrician, machinist, engineer or carpenter. Skilled laborers, should therefore be found to be high on S. Accountants, cashiers, bank tellers, sales clerks and the like are supposed to be favored by good arithmetic ability and should thus be high on Number ability (N). People who run their own business or belong to the managerial category would be expected to have an education and skills somewhere in between the professional and clerical groups and would thus be expected to be high on some attributes common to both.

[1] The data for this study were collected as part of the author's Ph.D. dissertation while a graduate student at University of Washington, Seattle. Financial assistance for processing the data was given by Science Research Associates, Chicago, and is gratefully acknowledged.

Method

A great many personality and other variables are involved in determining specific job choice and their control would be extremely difficult. It was therefore decided to concentrate upon a more general differentiation into 10 major occupational headings as used in the reports of the United States Census Bureau. Since the counseling use of the PMA is usually at the high school level, four of these major occupational classifications were selected for study as they are probably the most important ones being considered in a great majority of cases. These are: (*a*) professional and semiprofessional, (*b*) managerial and proprietary, (*c*) sales and clerical, and (*d*) skilled labor. In order to avoid artifacts introduced by transient or enforced job choice or possible PMA sex differences, only male *S*s were used. Since we are interested in stable occupational choice no *S* was to be included if he reported a change in his occupation or job specification over the past five years.

As part of another investigation, data were available on the PMA scores and the occupational status of a sample of 500 adult *S*s (**3**). Since age changes on the PMAs over the adult age range are known to be substantial, these were controlled experimentally by matching for age over the four occupational levels. From a pool of 172 *S*s who met the initial criteria for inclusion it was thus possible to match 20 sets of *S*s, or a total of 80 *S*s. These ranged in age from 26 to 65 years with a mean age of 45.5 years.

The S.R.A. Primary Mental Abilities test, intermediate form, was given to each *S* and was administered in group sessions using the instructions given in the examiner's manual. All raw scores were converted to standard scores with means of 50 and standard deviations of 10 by use of the norms available for the total sample of 500 adult *S*s. The reported mean scores are therefore directly comparable for the different mental abilities.

Results

The first step in the analysis of the test data was to compute means and standard deviations which are given in Table 1. The analysis of variance was then employed for a formal test of the null hypothesis with respect to over-all differences between the different PMAs and between occupational levels. The analysis for the total sample is presented in Table 2 and uses methods suggested by Edwards (**1**). The test for the difference between occupational levels is based on independent observations and thus uses within level variance as its error term. The test for the over-all differences between PMAs however, requires adjustment for correlation between the mental abilities. The pooled interaction of individuals and PMAs is therefore the correct error term for this test.

Inspection of Table 2 shows that *F* ratios for the variance associated with differences among PMAs as well as between occupational levels were found to be significant at the .001 level of confidence and the null hypothesis was therefore rejected. The interaction between PMAs and occupational levels, however, was not significant. These findings suggest that there are significant differences in over-all intellectual level between the different occupational groups as well as significant differences between scores on different abilities for most individuals. The lack of systematic interaction, however, indicates that specific PMA profile patterns are not a function of occupational level. It appears then that profile elevation, i.e. level of intelligence as estimated by the total PMA test, rather than profile pattern should be considered as the significant variable for predicting future occupational level.

TABLE 1

MEANS AND STANDARD DEVIATIONS ON THE PRIMARY MENTAL ABILITIES FOR DIFFERENT OCCUPATIONAL LEVELS

(N = 20 in each level)

	Skilled labor		Clerical & sales		Managerial & propr.		Professional & semi-prof.	
Verbal-meaning	44.0	9.5	51.5	7.4	50.9	10.2	55.1	5.8
Space	51.3	10.0	56.7	11.8	54.6	10.2	55.2	9.6
Reasoning	44.2	7.1	50.4	7.8	49.5	9.8	53.0	8.1
Number	48.3	8.7	54.5	8.6	55.2	7.8	57.7	10.4
Word-fluency	44.1	8.7	54.8	9.7	47.4	6.8	53.2	9.8

TABLE 2

ANALYSIS OF VARIANCE FOR THE COMBINED SAMPLE TESTING THE NULL HYPOTHESIS WITH RESPECT TO DIFFERENCES BETWEEN PRIMARY MENTAL ABILITIES AND BETWEEN LEVELS OF OCCUPATION

(N = 80; 5 scores for each S)

Source of variance	Sum of squares	*df*	Mean square	*F* ratio
Between levels	4,249.04	3	1,416.35	6.78*
Between individuals in level	15,877.56	76	208.92	
Total between	20,126.60	79		
Between PMAs	1,912.40	4	478.10	9.26*
Interaction: levels × PMAs	575.56	12	47.96	..
Pooled interaction: individuals × PMAs	15,688.44	304	51.61	
Total within	18,176.40	320		
Total variance	38,302.00	399		

* Significant at or beyond the .001 level of confidence.

The above analysis does not rule out the possibility that a given mental ability will tend to discriminate between different occupational levels while others do not. It is also possible that differences between Mental Abilities occur only in certain but not all of the occupational levels studied. To clarify these problems further analyses of variance were made for each separate occupational level and also for each of the different Mental Abilities.

The results shown in Table 3 indicate that there are indeed significant differences between the different PMA mean scores within both the "skilled labor" and "managerial" levels. Referring back to Table 1 it may be seen that for the "skilled labor" level Space is high, while low performance is found on all the verbal skills (V, R, and W). In the "managerial" group high scores are found to be Space and Number while this group is also low on V, R, and W. It is worthy of note that these patterns obviously overlap, explaining why the interaction between occupational level and abilities cannot be significant and why profile elevation turns out to be the significant discriminator.

Table 4 gives the results of the analysis of variance for the Primary Mental Abilities with respect to differences among occupational levels on each separate

TABLE 3

Analysis of Variance for the Differences Between Primary Mental Abilities in Each Separate Occupational Level, Adjusted for the Effect of Correlation Within Individuals
(N = 20; 5 scores for each individual)

	Between abilities		Within individuals		Residual error
	MS	*F*	*MS*	*F*	
Skilled labor	222.31	6.59*	284.22	7.52*	33.79
Sales & clerical	142.69	2.40	181.06	3.05*	59.43
Managerial	207.03	3.77*	206.58	3.67*	56.28
Professional	72.45	1.16	163.81	2.63*	62.36

* Significant at or above the .01 level of confidence.

ability. Verbal-meaning, Reasoning, and Word-fluency are found to differ significantly between occupational levels but Space and Number apparently fail to discriminate. Examination of the appropriate means shows high performance on Verbal-meaning and Reasoning for the professional group, low performance for the skilled laborers, and about equal and intermediate performance for the managerial and clerical groups. On Word-fluency the clerical and professional groups are about equal and high, while the managerial and skilled labor groups are low.

TABLE 4

Analysis of Variance for the Differences Between Occupational Levels on Each Separate Primary Mental Ability Adjusted for the Effect of Correlation Due to Matching for Age of *Ss*
(N = 80)

	Between occupational levels		Between matched individuals		Residual error
	MS	*F*	*MS*	*F*	
Verbal-meaning	604.45	11.99*	98.91	1.96	45.74
Space	108.41	..	112.80	..	106.38
Reasoning	277.15	4.57*	119.18	2.12	56.16
Number	318.18	3.94	95.97	..	80.64
Word-fluency	500.25	7.45*	100.09	1.49	67.04

* Significant at or beyond the .01 level of confidence. Trivial F ratios are omitted.

Another interesting analysis can be made by inspecting the standard deviations presented in Table 1. Since the population standard deviation has arbitrarily been assigned to be 10, the standard deviation for any subgroup would be expected to be significantly lower on any variable which tends to discriminate the subgroup from the total group. A low standard deviation would thus indicate that this is a variable on which the subgroup is more homogenous than the general population. Such increased homogeneity was found for the professional group on Verbal-meaning and for the managerial group on Word-fluency. Inspection of the range of standard deviations among the occupational levels gives further indications why some of the abilities fail to discriminate between levels.

Summary

Scores on the intermediate form of the S.R.A. Primary Mental Abilities test were examined for a stratified sample of male *Ss* from four occupational levels to test the hypothesis that differential performance on this test is useful in predicting future occupational placement. Several hypotheses frequently used in counseling on the basis of the PMA are presented and relevant evidence concerning the

PMA patterns of adults who have made permanent occupational choices is given.

The results of an analysis of variance yielded significant differences between the over-all ability for different occupational levels and also between different abilities. The interaction between occupational level and individual mental abilities, however, was not significant.

Significant differences were also found between abilities within the "skilled labor" and "managerial" groups. Analysis of the individual mental abilities showed significant differences between the mean scores for different occupational groups on Verbal-meaning, Reasoning, and Word-fluency.

It should be pointed out that the present study was concerned only with occupational levels. Pattern analysis of the PMA might therefore still be helpful for predicting success in a specific occupation. On the basis of the present findings, however it must be concluded that profile elevation (or general intellectual level) is of greater importance than profile pattern in predicting vocational choice.

REFERENCES

1. Edwards, A. L. *Experimental design in psychological research.* New York: Rinehart, 1950.
2. Mouly, G. M., & Robinson, G. M. A study of the United States Employment Service General Aptitude Test battery, B-1001. Univer. Minnesota, 1949.
3. Schaie, K. W. Rigidity-flexibility and intelligence: A cross-sectional study of the adult life span from 20 to 70. *Psychol. Monogr.*, 1958, 72, No. 9 (Whole No. 462).
4. Shaw, D. C. A study of the relationships between Thurstone's Primary Mental Abilities and high school achievement. *J. educ. Psychol.*, 1949, **40,** 239–249.
5. Thurstone, L. L., & Thurstone, T. G. Examiner Manual for the SRA Primary Mental Abilities Test, Intermediate Form. Chicago: Science Research Associates, 1949.

Received February 1, 1958

Journal of Educational Psychology
Vol. 49, No. 6, 1958

EFFECT OF INSTRUCTIONS ON FREE ASSOCIATION

ROSCOE A. BOYER
University of Mississippi

CHARLES F. ELTON
University of Kentucky

Although many investigators (**2, 3, 4, 8, 9, 10**) have demonstrated repeatedly that the counselor or test examiner may either deliberately or inadvertently structure the stimulus field, a review of the literature indicates that minimal research has been done regarding some of the characteristics of such influence. Only recently did Bordin (**1**) suggest the theoretical implications of the ambiguity structuredness variable in the counseling process.

It was the purpose of this study to investigate temporal effects and the verbal responses per se resulting from structuring the instructions regarding what would be appropriate responses using the free association technique. According to Bordin, if a group of "minimally anxious" *S*s were used, it could be deduced from his theoretical approach that the suggestions of appropriate responses to some words for these subjects would not influence the responses made to subsequent words. This study attempts to investigate this hypothesis.

A secondary purpose was to determine if there were regional differences occurring in the free associating technique.

Procedure

The *S*s were 401 college students attending the University of Mississippi during 1955–56 school year. Of these, 120 were enrolled in sophomore and junior year education courses, 130 in sophomore year psychology courses, and the remainder in economics, engineering and undergraduate courses in statistics. No sex differentiation was made. The *S*s were divided into three groups according to instructions given the students on the Kent-Rosanoff Free Association Test (**6**). An attempt was made to have equal representation of students from the various type classes in each of the three groups.

Group I. The instructions given by Russell and Jenkins (**7**) for administering the Kent–Rosanoff Free Association Test were used and are as follows with the exception that Mississippi was substituted for Minnesota:

> This is one of the studies in verbal behavior being done at Mississippi. This particular experiment is on free association.
>
> Please write your name on the outside of the paper passed to you. You can ignore the place for your name on the other side.
>
> When you open these sheets, you will see a list of 100 stimulus words. After each word write the first word that it makes you think of. Start with the first word; look at it; write the word it makes you think of; then go on to the next word.
>
> Use only a single word for each response.
>
> Do not skip any words.
>
> Work rapidly until you have finished all 100 words.
>
> When you are through, turn your paper over and write on the back the letter that appears on the board at that time.
>
> Are there any questions?
>
> Ready. Go.

The following additional section appeared after their third paragraph:

For Example, Your Responses to the First Words Might be as Follows:

No.	Stimulus	Response
1	Table	Chair
2	Dark	Light
3	Music	Song
4	Sickness	Health
5	Man	Woman

These example response words were those listed by Russell and Jenkins as the most common responses to the respective stimulus words. This group consisted of 133 students and will be referred to hereafter as the positively structured group.

Group II. The instructions to this group were the same as those for Group I except that the given "response word" examples had a frequency of 10 in 1031 samples as listed by Russell and Jenkins; therefore, these were considered uncommon or atypical responses to the respective stimulus words.

These response words were:

No.	Stimulus	Response
1	Table	Eat
2	Dark	White
3	Music	Dance
4	Sickness	Bad
5	Man	Mouse

This group consisted of 133 students and will be referred to hereafter as the negatively structured group.

Group III. The instructions given to this group were identical to those given by Russell and Jenkins; i.e., no response example was given. This group served as the control group and consisted of 135 students.

The above instruction for the three respective groups appeared on the first page of a three-page mimeographed test booklet. The second and third pages followed the form given by Russell and Jenkins (7).

In the present study the same procedure as that described by Russell and Jenkins was followed (7). After the students had been working on the test for four minutes, the letter A was printed on the blackboard. Every 30 seconds thereafter a new letter was substituted in alphabetical sequence. When the students had completed the test, they recorded on the back of the test booklet the letter appearing on the board at that time. Consequently a rough index of time necessary for each student to complete the test could then be obtained.

Following the administration of the tests, response frequencies were tabulated for each stimulus word for each of the three groups. The frequencies were then converted into percentages and tests of significances were made according to the procedure suggested by Lawshe and Baker (5).

A difference in percentage ratio was used to investigate the influence of the suggested answers. The formula used was $(E_1 - C)/C$, in which E_1 was the percentage of responses in the experimental group and C was the percentage of responses in the control group. The value of this ratio lies in offering a convenient way of showing how the values of the experimental groups converged with that of the control group.

Throughout this article, references to the expression, "most common response word," pertain to the word that was listed by Russell and Jenkins as the word having the highest response frequency for each of the 100 stimulus words. Unless otherwise indicated, all statistical analyses in this article are based upon the frequencies of these 100 most common response words. No attempt was made to evaluate differences of other response words.

Results and Discussion

Temporal effects. An analysis of variance was computed for the length of time required by the three groups to complete the tests. The *F* ratio was found to be 11.14 which was significant beyond the one per cent level. Consequently, these data were then examined for *t* values and the means and standard deviations are given in Table 1.

It was revealed that Groups I and II did not vary significantly from each other, but the differences were significant at the one per cent level between Groups I and III and between Groups II and III in

TABLE 1

Time in Seconds Required to Complete Test

Group	*N*	Mean	*SD*
I	133	536	139
II	133	511	121
III	135	473	121

time required to complete the tests. These data indicate that with the type of suggestion given to Groups I and II on a free association test, whether it be a common or uncommon answer, the time required to respond will be increased. It is to be noted that the most common or normal suggestions to a "normal" population resulted in the longest response time. The authors have no suggestion as to why this behavior occurred but are now in the process of attempting to duplicate this behavior with a different population and testing the hypothesis that differences in response time will disappear when certain variables are controlled.

Detailed examination of most common responses given to first 10 words. After all response words to the respective stimulus words had been tabulated, the frequencies were converted into percentages. In order to examine the influence of suggestion, a difference in percentage ratio was used and the data for the first 10 stimulus words are given in Table 2.

TABLE 2

Percentage of *S*s in Each Group Giving Most Common Response Word to Each of the First 10 Stimulus Words and the Difference Ratio Between University of Mississippi Groups

Stimulus	Response	Percentage of group responses				Difference ratio	
		I[a]	II[b]	III[c]	Minn,[d]	$\frac{\text{I} - \text{III}}{\text{III}}$	$\frac{\text{II} - \text{III}}{\text{III}}$
1. Table	Chair	94	28	81	84	.16	−.65
2. Dark	Light	87	50	71	83	.23	−.30
3. Music	Song/s	58	11	15	18	2.86	−.27
4. Sickness	Health	67	24	30	38	1.23	−.20
5. Man	Woman/en	92	66	81	77	.14	−.19
6. Deep	Shallow	60	47	44	32	.36	.07
7. Soft	Hard	64	56	55	45	.16	.02
8. Eating	Food	34	24	35	39	−.03	−.31
9. Mountain	Hill/s	44	44	40	27	.10	.10
10. House	Home	33	29	28	25	.18	.04

[a] *N* for group I = 133.
[b] *N* for group II = 133.
[c] *N* for group III = 135.
[d] Percentages are based on Russell and Jenkins data.

It should be recalled that differences are based upon but one word in each response set; that is, the most common response word as given by Russell and Jenkins. These data in Table 2 revealed that by the time the students reached the sixth stimulus word, for all practical purposes, the influence of the suggested words had been dissipated. This trend was more consistent and stable in the negatively structured group (Group II) than in the positively structured group (Group I). The difference in response frequency for the most common response words between Groups I and II and between Groups I and III was significant at the one per cent level for the first four stimulus words.

Also, there was found a significant difference at the one per cent level between Groups II and III for responses given to stimulus words [1] Table and [2] Dark. A possible explanation of the lack of significance between Groups II and III in response frequency to stimulus words [3] Music and [4] Sickness is that the habit strength for the most common response might be considered relatively weak. According to Russell and Jenkins the most common response for the stimulus word Music occurs with a frequency of 18 per cent and the frequency for Sickness is 38 per cent (7). The fact that percentage differences are significant between Groups I and II and Groups I and III for these same words may be a function of the suggestions given in the instructions and/or the lack of a strong competing response habit strength. The authors are now investigating this possibility by ranking the words in the Kent-Rosanoff word list according to the response strength of each stimulus word and repeating this study.

The explanation for the significant difference in percentage response for the stimulus word [5] Man between Groups I and II and the absence of a significant difference between Groups I and III and Groups II and III is more difficult. It may be that negative suggestion dissipates more rapidly than positive suggestion and this could have been operating to produce such an effect. More likely is the fact that rate of dissipation is confounded with the problem of unequal response habit strengths among first five words.

Examination of the remaining 90 most common response words. As indicated above, the instructions did not influence the response beyond the fifth word. The only significant differences that were found between any two of the three groups were for the stimulus words [20] Chair, [23] Woman, [59] Health, and [88] Heavy. Although these differences were significant at the one per cent level, an a priori explanation would be that they have occurred by chance and were due neither to the instructions nor to the samples that were used.

Regional differences. Regional differences were examined by comparing the response frequencies of college students at the University of Minnesota with the response frequencies made by the control group students at the University of Mississippi. Only one response word was found to be significantly different: i.e., for stimulus [24] Cold, at the University of Minnesota, 34 per cent gave the response Hot, whereas 60 per cent of the college students at the University of Mississippi gave that response.

SUMMARY

Four hundred and one undergraduate college students were divided into three groups for administration of the Kent-Rosanoff Word Association Test, under the following conditions: The first group of students was given five examples of common responses to the stimulus words; a second group of students was given five examples of uncommon responses or responses occurring approximately one per cent of the time; and a third group of students was given no example. The resulting response frequencies were compared. There was no apparent difference in responses after the sixth word among those students who were given common or "normal" response examples, those given atypical responses, and those given no example of responses to the stimulus words. No differences were found between responses given by college students at the University of Minnesota and college students at the University of Mississippi. This research suggests that the influence of any response instructions given to a normal population on a free association word test will be rapidly dissipated.

REFERENCES

1. Bordin, E. S. Ambiguity as a therapeutic variable. *J. consult. Psychol.*, 1956, **19**, 9-15.
2. Cameron, N., & Margaret, A. *Behavior pathology*. Boston: Houghton Mifflin, 1951.
3. Gibby, R. G., Miller, D. D., & Walker, E. L. The examiner's influence on the Rorschach protocol. *J. consult. Psychol.*, 1953. **17**, 425–428.
4. Greenspoon, J. The effect of verbal and nonverbal stimuli on the frequency of members of two response classes. Unpublished doctoral dissertation, Indiana Univer. 1950.
5. Lawshe, C. H., & Baker, P. C. Three aids in the valuations of the significance of the difference between percentages. *Educ. psychol. Measmt.*, 1950, **10**, 263–270.
6. Rosanoff, A. J. *Manual of psychiatry*. New York: Wiley, 1927. Pp. 546–604.
7. Russell, W. A., & Jenkins, J. J. The complete Minnesota norms for responses to 100 words from the Kent-Rosanoff Word Association Test, *Studies on the role of language in behavior. Technical Report No. 11.* August, 1954. Univer. Minnesota Dep. Psychol.
8. Sacks, E. L. Intelligence scores as a function of experimentally established social relationships between child and examiner. *J. abnorm. soc. Psychol.*, 1952, **47**, 354–358.
9. Schachtel, E. G. Subjective definitions of the Rorschach test situation and their effect on test performance. *Psychiatry*, 1945, **8**, 419–448.
10. Wickes, T. A. Jr. Examiner influence in testing situation. *J. consult. Psychol.*, 1956, **20**, 23–26.

Received March 31, 1958.

Journal of Educational Psychology
Vol. 49, No. 6, 1958

COLLEGE STUDENT STEREOTYPES OF THE PERSONALITY TRAITS OF RESEARCH SCIENTISTS

A. W. BENDIG AND PETER T. HOUNTRAS
University of Pittsburgh

The extent of public support for scientific research and education is dependent upon the attitudes toward science and scientists which prevail in the culture. The development of these attitudes begins early in the elementary school (**2**). These attitudes are solidified by the time students reach the secondary school level (**6, 8**) where they influence the choice of a future career (**7**). The attitude of the public toward the current Man-Into-Space program is at least partially influenced by a general attitude of both respect for and a fear of the influence of scientific advances upon our society. Stated somewhat differently, this ambivalent feeling toward the research scientist means that he is simultaneously a "different" and perhaps slightly dangerous individual, but also a necessary and even useful member of our society. These attitudes would appear to be particularly significant as far as elementary and secondary school teachers are concerned since they are in daily contact with potential future scientists during the period when these attitudes are developing.

An indirect and somewhat disguised approach to *S*s' attitudes toward scientists may be through a study of the consistency with which *S*s attribute a syndrome of personality characteristics to the average scientist. It is clear that stereotypes of the personality traits of people in various occupations do exist among college students (**1, 9**) and identification of the specific traits that *S*s believe distinguish the scientist from people in other occupations may provide an insight into their attitudes toward the scientist and permit the subsequent development of a relatively simple and objective assessment device to measure these attitudes. Terman's (**10**) investigation of intellectual and interest differences among four occupational groups, scientists, engineers, lawyers, and businessmen, suggests that comparisons among the personality traits attributed to these occupations might provide evidence as to the students' stereotype of the personality of the research scientist.

Procedure

Traits. An original list of approximately 100 personality trait names was compiled from several published sources (**1, 8, 9**). Subsequently a list of 60 traits was selected on the basis of two criteria: (*a*) 30 traits that appeared on an a priori basis to be socially desirable and 30 traits judged to be socially undesirable, and (*b*) traits were selected that appeared representative of many significant dimensions of behavior including work habits, intellectual characteristics, and both the social and nonsocial aspects of personality. These criteria were employed (*a*) to minimize response bias in the subsequent ratings and to include a wide range of social desirability in the selected personality characteristics, and (*b*) to insure as far as possible an adequate sampling of many different areas of behavior. The trait names finally selected were: accurate, calm, clumsy, fearful, considerate, meddlesome, intellectual, economical, democratic, inept, egotistical, cruel, logical, mature, unsystematic, pessimistic, friendly, sarcastic, studious, alert, kind, disorganized, timid, critical, orderly, responsible, incompetent, impulsive, tactful, annoying, precise, sincere, humorous, unimaginative, reckless, irritable, persistent, stable, sloppy, nervous, sympathetic, shy, thorough, self-confident, generous, unproductive, miserly, fault-finding, industrious, dependable, inefficient, moody, tolerant, argumentative, capable, unreliable, rigid, poised, lonely, charming.

Forms. Four occupational titles were selected (research scientist, engineer, lawyer, businessman), similar to the comparison groups used by Terman (**10**), and six forms were prepared, one form for each of the possible combinations of two of the four occupations. On each form the *S* was requested to compare the average person in one (rated) occupation with the average person in another (reference) occupation and to make a judgment as to whether the first person would be most likely to have more, less, or an equal amount of each of the 60 traits than the person in the second (reference) occupation. For example, the significant portions of the instructions for Form C were:

> We all know that a person's interests, abilities, attitudes, and personality characteristics determine to a large extent what occupation he or she will select. For example, the average *research scientist* has more or less of certain traits than does the average *businessman,* although on other traits these two people will have the same amount of these particular traits. We are asking you to identify which of these traits distinguish the *research scientist* from the *businessman* and which traits they have in common.
>
> Below is a list of 60 traits to be identified.... Please indicate on your answer sheet your judgment for each of the traits using the following marking system:
>
> Column A: the average *research scientist* has *more* of this trait than the average *businessman.*
>
> Column B: both the average *research scientist* and the average *businessman* have about the *same amount* of this trait.
>
> Column C: the average *research scientist* has *less* of this trait than the average *businessman.*

The combinations of occupations used on the forms are as follows: (*a*) Research Scientist vs. Engineer; (*b*) Research Scientist vs. Lawyer; (*c*) Research Scientist vs. Businessman; (*d*) Engineer vs. Lawyer; (*e*) Engineer vs. Businessman, and (*f*) Lawyer vs. Businessman.

A seventh form (Form G) was constructed that requested the *S*s to rate each of the 60 trait names on a five-point scale of social desirability. No reference was made on this form to specific occupations, but the *S*s were told that we were trying to obtain relative measures of the social desirability of a large number of personality traits. This last form was used as a check on our original dichotomization of the traits into socially desirable and undesirable groups.

Subjects. Form G was administered to 54 *S*s (18 men and 36 women) enrolled in two sections of introductory educational psychology. Forms A through F were randomly distributed to 154 *S*s in four other sections of the same course, each *S* receiving only one form. Sixteen *S*s were discarded from this second group to insure that equal numbers of men and women *S*s responded to each form. The discarding of *S*s from each form-sex subgroup was random and the final group consisted of 138 *S*s with 23 *S*s (8 men and 15 women) recording their judgments on each of the six forms. The *S*s were sophomore pre-education students who are required to take this course prior to admission to the School of Education.

Results

The mean social desirability rating of each trait by the 54 *S*s who received Form G was computed and no overlap was found in mean ratings between the 30 traits that had been a priori selected as socially desirable and the 30 traits selected as socially undesirable. Consequently, the original grouping of the items into these two classes was retained in subsequent analyses.

The answer sheets of the 138 *S*s who responded to Forms A through F provided four separate scores: the number

TABLE 1

Analyses of Variance of Four Personality Trait Scores Distinguishing Between Combinations of Occupations

Source of Variation	*df*	Desirable-More		Desirable-Less		Undesirable-More		Undesirable-Less	
		Mean Square	*F*	Mean Square	*F*	Mean Square	*F*	Mean Square	*F*
Sex (Sx)	1	5.22	.29	38.74	2.77	32.36	2.13	36.27	2.05
Forms (F)	5	147.60	8.19*	138.06	9.87*	18.01	1.19	11.92	.67
Sx × F	5	16.67	.92	7.61	.54	17.46	1.15	24.53	1.39
Within	126	18.03		13.99		15.19		16.67	

* $P > .001$

of desirable traits the rated occupation had more of (Desirable-More), the number of desirable traits the rated occupation had less of (Desirable-Less), the number of undesirable traits the rated occupation had more of (Undesirable-More), and the number of undesirable traits the rated occupation had less of (Undesirable-Less). Each of these four scores was then separately subjected to a two-criterion (sex and forms) analysis of variance. The results are reported in Table 1. Both the Desirable-More and Desirable-Less scores discriminated among the six forms at the .001 level of confidence, but neither the Undesirable-More nor the Undesirable-Less scores gave any evidence of significant differences among the forms. No statistically significant (.05 level) sex differences were found in any of the analyses. Apparently the *Ss* could consistently discriminate differences among the pairs of occupations with respect to the presence or absence of desirable traits shown by the average individual in the four occupations, but did not discriminate among the occupations as to undesirable personality traits. This suggests that stereotypes concerning occupational personalities involve the presence or absence of socially desirable traits only.

TABLE 2

Mean Numbers of Desirable Traits Attributed to Four Occupations in Six Combinations ($n = 23, N = 138$)

Form	Occupation Rated	Reference Occupation	Desirable Traits		
			More	Less	Difference
A	Scientist	Engineer	9.7	5.7	4.0
B	Scientist	Lawyer	6.7	7.3	−0.6
C	Scientist	Businessman	10.0	6.8	3.2
D	Engineer	Lawyer	5.8	10.7	4.9
E	Engineer	Businessman	7.0	6.9	0.1
F	Lawyer	Businessman	12.4	3.1	9.3

The mean number of desirable traits attributed to each pair of occupations can be found in Table 2. The difference score between the "more" and "less" means can be viewed, in a sense, as a "favorability" score for the rated occupation when compared with the reference occupation. Both the sum of the "more" and "less" scores and the difference between these two scores were subjected to two-criterion (sex and forms) analyses of variance. The sums did not discriminate between the forms ($F = 1.23$) while the difference score showed form differences that were significant at the .001 level ($F = 14.75$). No sex differences were found in either analysis. Duncan's method was used to test the significance of the differences among the difference score means (**4, 5,** pp. 26–29)

TABLE 3

SINGLE PERSONALITY TRAITS MOST CONSISTENTLY ATTRIBUTED TO SIX COMBINATIONS OF OCCUPATIONS

Occupations Compared	The First Occupation Is		Occupations Compared	The First Occupation Is	
	More	Less		More	Less
Research Scientist vs. Engineer	Persistent Studious Intellectual Alert Thorough	Economical	Engineer vs. Lawyer	Precise	Poised Charming Self-confident Tactful Alert
Research Scientist vs. Lawyer	Precise	Charming Poised Humorous Self-confident Friendly	Engineer vs. Businessman	Precise Accurate Studious Thorough	Tactful Humorous Poised
Research Scientist vs. Businessman	Thorough Studious Precise Intellectual Orderly Persistent Accurate Logical	Charming Tactful Friendly Humorous Economical	Lawyer vs. Businessman	Studious Intellectual Poised Precise Thorough Logical Persistent Tactful Tolerant	Economical

and it was found that the difference score means fell into four groups. The difference mean of Form D was significantly (.05 level) larger than the difference means of Forms A and B. Forms A and B were significantly different from Forms C and E, and the mean difference score of Form F was significantly lower than the mean difference scores of Forms C and E. The differences in means between Forms A and B and also between Forms C and E were not significant. The research scientist and the lawyer were quite similar in difference ("favorability") scores as were the engineer and businessman. However, the scientist-lawyer pair of occupations were quite distinct from the engineer-businessman in mean difference scores.

Analyses were performed for the 30 desirable traits on each form to identify the single traits that distinguished between each pair of occupations. The "more" and "less" percentages ($N = 23$) were computed for each trait on each form and if either percentage was greater than 57 per cent the trait was selected as being a discriminating item. The results of these individual trait analyses can be found in Table 3. It seems apparent that stereotypes exist for all four occupations. The scientist and lawyer appear similar in the more intellectual traits, but the scientist lacks the warm social graces that characterize the lawyer stereotype. The engineer is a junior edition of the scientist while the businessman lacks the intellectual qualities of the lawyer, but shares many of his social traits.

Since our basic interest was in the

stereotype of the research scientist, the traits that most consistently discriminated the scientist from the other three occupations are given in Table 4. Again the same stereotype pattern appears: the scientist, along with the lawyer, has more of the socially desirable intellectual and work habit traits than does the engineer and the businessman, while the scientist, like the engineer, has less of the social graces than does the lawyer and businessman. The 12 traits listed in Table 4 appear to constitute the core stereotype that the *S*s had regarding the research scientist.

Discussion

The evidence that our college *S*s consistently attribute certain personality traits to the occupations of research scientist, engineer, lawyer, and businessman confirms the results of other studies (**1, 9**) where different occupations and different methodologies were used. The finding of most general interest was that the *S*s discriminated among the occupations only for socially desirable traits and not for socially undesirable traits. Whether this implies reluctance on the part of the *S*s to say that one occupation has more of these undesirable traits, or hesitation in attributing relative freedom from undesirability traits to the paired occupation, cannot be determined from our data. The implications for research where the *S* is required to attribute personality traits to himself and also to other people are obvious.

The analysis of individual traits comprising the stereotypes of the personalities of these four occupations suggests that the occupations were discriminated along two relatively independent dimensions. The traits can be grouped into (*a*) those referring to intellectual and work habits characteristics and (*b*) those related to social personality traits that arise primarily in interpersonal relations. Research scientists are viewed as being high (having more of the traits) on the intellectual dimension and as being low (having fewer of the traits) on the social axis. Engineers are not as intellectual as scientists, but members of both professions are equally lacking in social graces. The lawyer is equally high on both axes, while the businessman is low on the intellectual dimension and high on social traits. Whether or not other occupations can be located within this two-dimensional system, or whether other dimensions would have to be added are questions for further study.

TABLE 4

Percentages of Subjects Saying that Certain Socially Desirable Traits Distinguish the Research Scientist from Men in Other Occupations

Research Scientist is *More*	Than the Average		
	Businessman	Engineer	Lawyer
Intellectual	70	70	26
Logical	61	52	26
Orderly	70	48	48
Persistent	70	91	30
Precise	87	39	78
Studious	91	78	43
Thorough	96	65	39

Research Scientist is *Less*	Than the Average		
	Businessman	Engineer	Lawyer
Charming	70	34	83
Friendly	65	39	57
Humorous	61	39	70
Poised	52	30	74
Self-confident	35	48	61

The 12 traits listed in Table 4 appear to offer a possibility of developing a short objective scale for measuring the extent of the stereotype of the research scientist held by individual *S*s. The same procedure used in this study could be repeated by administering Forms A, B, or C to *S*s and scoring each *S* as to how many of the first seven traits the *S* says the scientist has more of and how many of the last five traits he indicates the scientist has less of. Such a short 12-item scale may lack sufficient reliability for research purposes, but, if reliable, would offer a method of quantifying this aspect of attitudes for further research.

It is particularly interesting to note that the *S*s displaying this stereotype of the research scientist were pre-education students who, in a few years, will be teaching children and adolescents from whom the next generation of scientists must be recruited. If this stereotype continues to be transmitted from teacher to student, the problem of interesting high school students in scientific careers will remain with us.

Summary

Pre-education college *S*s ($N = 138$) were asked to compare two of four occupations (research scientist, engineer, lawyer, businessman) as to whether the average members of the paired occupations would have more, less, or an equal amount of each of 60 personality traits. Equal numbers of *S*s ($N = 23$) responded to each of the six possible pairs of occupations. The traits were evenly dichotomized into socially desirable and socially undesirable groups on the basis of an a priori selection which was validated by having another group of *S*s ($N = 54$) rate the traits for social desirability. The number of socially desirable traits attributed to each occupation discriminated among the occupations (.001 level), but the socially undesirable traits did not. Analyses of the 30-individual socially desirable traits indicated that the *S*s viewed the scientist and lawyer as having more of the intellectual traits while the lawyer and the businessman were perceived as having more of the desirable interpersonal traits. Both the scientist and engineer have less of the interpersonal traits and the businessman has few of the intellectual traits. The most consistent stereotype in this study regarded the research scientist, when compared with the other three occupations, as being more intellectual, logical, orderly, persistent, precise, studious, thorough, and also as being less charming, friendly, humorous, poised, and self-confident.

REFERENCES

1. Borg, W. R. The effect of personality and contact upon a personality stereotype. *J. educ. Res.*, 1955, **49,** 289–294.
2. Brown, S. B. Science information and attitudes possessed by California elementary school pupils. *J. educ. Res.*, 1954, **47,** 551–554.
3. Cattell, R. B., & Drevdahl, J. E. A comparison of the personality profile (16 P.F.) of eminent researchers with that of eminent teachers and administrators, and of general population. *Brit. J. Psychol.*, 1955, **46,** 248–261.
4. Duncan, D. B. Multiple range and multiple *F* tests. *Biometrics*, 1955, **11,** 1–42.
5. Federer, W. T. *Experimental design.* New York: Macmillan, 1955.
6. Remmers, H. H., & Radler, D. H. *The American teenager.* Indianapolis: Bobbs-Merrill, 1957.
7. Roe, Anne. *The making of a scientist.* New York: Dodd, Mead, 1953.
8 Schlesinger, L. E. Public attitudes toward science: I. Attitudes of secondary school female students. *J. soc. Psychol.*, 1954, **40,** 211–218.
9. Secord, P. F., Bevan, W., Jr., & Dukes, W. F. Occupational and physiognomic stereotypes in the perception of photographs. *J. soc. Psychol.*, 1953, **37,** 261–270.
10. Terman, L. M. Are scientists different? *Sci. Amer.*, 1955, **192,** 25–29.

Received May 22, 1958.

Journal of Educational Psychology
Vol. 49, No. 6, 1958

AUDITORY ABILITIES AND ACHIEVEMENT IN SPELLING IN THE PRIMARY GRADES[1]

DAVID H. RUSSELL

University of California, Berkeley

School people are necessarily concerned with language habits and attitudes that develop in the primary grades. When these learnings involve spelling they may influence a child's achievement in written language not only in the beginning grades but throughout later years. In his first years in school the child utilizes language skills acquired in preschool years and may, therefore, rely heavily on auditory clues to words he must spell. There is some evidence that auditory abilities are more important for spelling in the lower grades than they are by the time the child reaches the seventh grade level of spelling ability **(11)**. Evidence of the close relationship between auditory and spelling abilities in the primary grades has been suggested by such investigators as Bradford **(1)** and Russell **(10)**. Bradford used a revised paper-and-pencil test of individual vowel and consonant sounds and blends of "regularly spelled" words and found considerable growth between the first and second grades. Russell found correlations between spelling and auditory tests ranging from the .20's to .80's in a second grade group. Typical of the investigations in the area is one by Rudisill **(9)** which found a correlation of .69 between spelling and phonic knowledge for a group of 315 third grade children in North Carolina. In another study of the effects of phonic training at the second-grade level, Zedler **(13)** found improvement, after 14 hours of instruction, in both spelling scores and speech-sound discrimination abilities.

Although spelling has usually been regarded as one of the simpler skills acquired in school, one with a heavy loading of associative learning, Horn **(7)** and others have shown just how intricate and complex the relationship between sounds and letters may be. The apparently plain injunction to combine phonetic analysis in spelling and reading activities is not so simple nor so direct as it seems. Horn, for example, has listed six types of evidence which must be considered in relating auditory characteristics of words to their spelling and has presented facts about three of them: (*a*) the variation in pronunciations of the "same" words, (*b*) the different ways in which the various English sounds are spelled, and (*c*) the ways children actually spell sounds in common words. He concludes, for example, that there is little justification for the claim that children can spell the words they can pronounce and therefore believes that direct teaching of the large number of irregularly phonetic words is inevitable.

Since many words must be taught directly in the lower grades, the question every teacher faces is that of how to get children to study the words efficiently. Shall this child be encouraged to rely on visual techniques? Does that child do better with auditory techniques, and if so, which ones should he use? The present study is concerned with identifying auditory techniques which a child is most likely to find useful at the primary-grade level.

Procedures

To explore further the relationships between auditory abilities and spelling, 97 children in the first three grades of an Oakland, California, school were tested. The numbers used were Grade I, 30;

[1] Barbara J. King and Gerald M. Meredith assisted in collecting and processing the data.

Grade II, 32; and Grade III, 35. Since some of the tests were not useful for children reading at the pre-primer level, the results in the study are obtained largely from a sample of 85 children with somewhat less than a third of the group in first grade. The children came largely from middle or lower-middle class homes. In the three grades the range in CA was from 6–8 to 10–2, in MA from 6–7 to 10–2, and in IQ from 83 to 119.

The following standardized tests were administered:

1. The Kuhlmann-Anderson Intelligence Tests
2. The California Achievement Test—Spelling
3. The Durrell-Sullivan Reading Capacity Test
4. The Gates Primary (and Advanced Primary) Reading Tests, Types I, II and III.

In addition, six tests of auditory discrimination were given to the children. Since not all of these have been published they are described briefly, with examples. The group tests were:

1. Caffrey-Russell Auditory Discrimination Test I Same-Different. The teacher reads pairs of words such as "shown-sewn," "style-style" and "mobbed-mopped" and the child marks whether they are the "Same" or "Different." (This test proved to be too easy for the group.)

2. Caffrey-Russell Auditory Discrimination III telling whether words are different in initial, middle, or final sounds. The children mark 1, 2, or 3 (corresponding to initial, middle, final) on an answer sheet for such pairs as "butter-buzzer," "pits-pitch," and "shoed-chewed."

3. Durrell Test of Hearing Sounds in Words. This test (**2**) consists of three subtests: (*a*) marking the printed word which has an initial sound the same as a word given orally. Example: The teacher says "top" and the children mark one letter of p, b, t, n, a. (*b*) Marking the printed word which has the same final or beginning sound as a word given orally. Example: The teacher says "happen" and the child marks one of hexameter, generation, and hydrogen. (*c*) The pupil draws a circle around all phonetic elements (such as letters, blends, phonograms) heard in a given word. For example the teacher says "blinding" and the child marks the following: ind r bl x t ing.

4. Durrell-Sullivan Reading Capacity Test. This is a test of comprehension of paragraphs read orally which might be called a listening or auditory test rather than a reading test. The eight paragraphs used were modified slightly from the Durrell-Sullivan Capacity Test so that raw scores were used in computation. Each paragraph was followed by five oral questions in which the child marked one of three possible answers.

Two tests were given individually. These were:

5. The Gates test of Giving Words with Stated Initial Sounds described in *The Improvement of Reading* (**3**) in which the *S* is asked to name three words which begin like each of three words suggested by the examiner.

6. Gates test of Giving Words with Stated Final Sounds which is described in the same source. The child is asked to say three words which rhyme with each of three words suggested by the examiner.

Results

Table 1 gives the zero-order correlations for the various tests with spelling scores. The table indicates that the reading tests as a group correlate more highly with spelling than the individual auditory tests but that the combined group or battery of auditory tests correlate with spelling as highly as the Gates reading tests. The table further suggests that, for this group, the best test of auditory abilities in relation to spelling is the Durrell test, composed of three subtests. Chronological age and mental age do not seem closely related to spelling ability. In general, the results suggest that rather complex auditory abilities involving sound recognition in various parts of a word are more closely related to spelling ability than is recognition of sounds of whole words as in same-different or rhyming tests. The close relationship of the Gates reading tests, especially in word recognition, to spelling tends to confirm an earlier finding (**10**). In addition to

TABLE 1

Correlations of Various Factors with Spelling Ability for 85 Children in Grades I, II, and III

Variable	Zero-order r With Spelling*
General	
1. Kuhlmann-Anderson Mental Age[a]	.31
2. Chronological Age	.17
Reading	
3. Gates Type I	.63
4. Gates Type III	.57
5. Gates Total	.65
Auditory	
6. Caffrey-Russell I	.22
7. Caffrey-Russell III	.51
8. Durrell Sounds	.66
9. Gates Initial Sounds	.29
10. Gates Rhyming	.22
11. Listening Comprehension (Durrell-Sullivan)	.33
12. Auditory Total (Items 6 to 11)	.66

* For $n = 85$, $r = .22$ significant at 5% level, $r = .28$ significant at 1% level.

[a] $n = 58$.

these calculations, the raw scores of the auditory tests were converted to standard scores, but the correlations computed gave about the same correlations as the raw scores.

Table 2 illustrates the use of the coefficient of multiple correlation to estimate the relationship to spelling of four combined auditory tests. The contribution to variance may be computed by multiplying the zero-order correlation by its standard partial regression coefficient. In this study the Doolittle Method was used in computing the standard partial coefficients as described in Walker and Lev (**12,** pp. 326–331). Once computed, the standard partial regression coefficients are inserted in the conventional formula for the multiple correlation, and a measure of contribution of multiple factors to a set of scores may be estimated. Table 2 illustrates that factors other than auditory abilities account for spelling achievement in the group tested but that the relationship of auditory abilities to spelling ability is a significant one.

Table 3 gives some further information

TABLE 2

Multiple Correlation and Contribution to Variance of Spelling by Selected Auditory Tests

Auditory Test	Zero-order r With Spelling	Cum. Mult. Correlation	Contribution to Variance
1. Durrell Sounds	.66	.66	35%
(1) plus 2. Caffrey - Russell III	.51	.69	13
(1) and (2) plus 3. Gates Initial Sound	.29	.71	3
(1) and (2) and (3) plus 4. Caffrey-Russell I	.22	.72*	1
Total Variance Accounted for	. . .	. . .	52%

* Significant at the 1% level of confidence.

TABLE 3

Intercorrelations of Auditory Tests

	I	II	III	IV	V	VI	Total Auditory Score
I		.27	.18	.47	.36	.23	.41
II	.27		.49	.42	.25	.26	.69
III	.18	.49		.42	.29	.28	.79
IV	.47	.42	.42		.43	.35	.65
V	.36	.25	.29	.43		.55	.48
VI	.23	.26	.28	.35	.55		.42

Note.—Code: I Caffrey-Russell Auditory Discrimination Test I
II Caffrey-Russell Auditory Discrimination Test III
III Durrell Sounds in Words Test
IV Listening Capacity Test (adapted from Durrell-Sullivan)
V Gates Giving Words With Same Initial Sound
VI Gates Giving Words that Rhyme

about the interrelationships of the auditory abilities involved in this study. It indicates that the combined score obtained on the battery of three Durrell tests is most closely related to the total auditory scores. As mentioned above, the Caffrey-Russell I test as administered was too easy for this group with a considerable number of top scores reducing the size of the correlations and the discrimination value of the test. It should also be noted that the Listening Capacity Test, which was a measure of comprehension of verbal materials, had a fairly high correlation with the other auditory perception tests.

Conclusions

This study of 85 children in the first three grades revealed that some auditory abilities are significantly related to spelling abilities at the one per cent level of confidence. It began an exploration of specific auditory abilities which are most closely related to spelling achievement and found that these were rather complex abilities involving word parts rather than whole words. A battery of three Durrell tests of word sounds and the Caffrey-Russell III test which involved recognition of likenesses in initial, middle or final positions were closely enough related to spelling ability to warrant further study. There is considerable evidence that a group of auditory abilities can be good predictors of spelling success in the primary grades but the constituents of this group must be studied more broadly and more exactly. The hypothesis that different children have quite different auditory abilities and therefore should be taught spelling, and possibly reading, with different kinds of auditory techniques also needs further testing.

In addition to the role of auditory abilities in spelling achievement, the investigation has confirmed earlier results of the close relationship between the Gates tests of primary reading and spelling ability at this age. On the other hand, the factors of chronological age and mental age within this fairly narrow age range are not significant factors in spelling ability.

The relationship of listening comprehension of oral paragraphs to the auditory and spelling tests is a puzzling one. Better measures of listening or auding ability are needed. If the test used in this study is a valid one it appears that the ability to listen to paragraphs with comprehension is not closely related to spelling ability ($r = .33$) but that it is fairly closely related to the combined auditory perception or discrimination scores ($r = .65$). This fact, and the close relationship of the spelling scores to the Gates word recognition test indicate the presence of other factors, possibly visual discrimination abilities, which were not considered in the present study.

The results further indicate the need of complete exploration of different kinds of phonetic and auditory abilities and their relations to spelling achievement. In addition to the six measures used in this study possible tests include other tests devised by Bradford (**1**), by Durrell (**2**), by Holmes (**6**), by Roswell-Chall (**8**) and others. The present study suggests that the simpler skills of detecting same-different word pairs or suggesting rhymes are not so closely related to spelling achievement as are more complex auditory abilities such as identifying sounds in various parts of words. Knowing when similar sounding syllables are alike and different, and knowing the various ways a syllable may be spelled once it has been recognized, make the apparently simple process of spelling more complex than it first seems.

Summary

The relation of scores on the six tests of auditory discrimination, sometimes labelled "phonetic skills," to scores on intelligence, spelling, and reading tests was determined for 85 children in the first three grades of a California school. The results

indicated that some verbal auditory skills are significantly related to both spelling and reading ability and that these abilities involved recognition of word parts rather than whole words. The relationship of listening comprehension of paragraphs to spelling scores was much lower. Considerable contribution to spelling variance was unaccounted for indicating the possibility that visual discrimination factors may be important in spelling or that a wider range of specific kinds of auditory skills should be tested probably in relation to both spelling and reading.

REFERENCES

1. Bradford, H. F. Oral-aural differentiation among basic speech sounds as a factor in spelling readiness. *Elem. Sch. J.*, 1954, **54**, 354–358.
2. Durrell, D. D. *Improving reading instruction.* Yonkers-on-Hudson: World Book, 1956.
3. Gates, A. I. A study of the role of visual perception, intelligence, and certain associative processes in reading and spelling. *J. educ. Psychol.*, 1926, **17**, 433–445.
4. Gates, A. I. *The improvement of reading.* (3rd ed.) New York: Macmillan, 1947.
5. Gates, A. I. & Chase, Esther, H. Methods and theories of learning to spell tested by studies of deaf children. *J. educ. Psychol.*, 1926, **17**, 289–300.
6. Holmes, J. A. Phonetic Association Test. (Multilith ed). Univer. California, Berkeley, 1953.
7. Horn, E. Phonetics and spelling. *Elem. Sch. J.*, 1957, **57**, 424–432.
8. *Roswell-Chall Diagnostic Reading Test of Word Analysis Skills.* New York: Essay Press, 1956.
9. Rudisill, Mabel F. Interrelations of functional phonic knowledge, reading, spelling and mental age. *Elem. Sch. J.*, 1957, **57**, 264–267.
10. Russell, D. H. A diagnostic study of spelling readiness. *J. educ. Res.*, 1943, **37**, 276–283.
11. Russell, D. H. A second study of characteristics of good and poor spellers. *J. educ. Psychol.*, 1955, **46**, 129–141.
12. Walker, Helen M., & Lev, J. *Statistical inference.* New York: Holt, 1953.
13. Zedler, Empress Y. Effect of phonic training on speech sound discrimination and spelling performance. *J. Speech Dis.*, 1956, **21**, 245–250.

Received June 2, 1958.

Journal of Educational Psychology
Vol. 49, No. 6, 1958

COMPARISON OF PRESCHOOL STANFORD-BINET AND SCHOOL-AGE WISC IQS[1]

FRANCES FUCHS SCHACHTER AND VIRGINIA APGAR

College of Physicians and Surgeons, Columbia University

In several studies, the Stanford-Binet, Form L, and the Wechsler Intelligence Scale for Children (WISC) were administered to the same children, at the same age, by the same examiner (**2, 5, 7, 9, 10, 11, 12, 13, 15**). The following results were obtained: (*a*) The median correlation between the Stanford-Binet and the WISC Full Scale IQ was .85. (*b*) Highest intertest correlations obtained for the WISC Full Scale, next highest for the WISC Verbal Scale, and lowest for the WISC Performance Scale (**2, 5, 7, 10, 11, 12, 13**). (*c*) Mean Stanford-Binet IQs were significantly higher than mean WISC IQs (**10, 11, 12, 13**). (*d*) Significantly greater intertest discrepancies occurred at the high IQ and low age levels (**10**).

Previous finding may require modification before application to the situation where retesting occurs at different ages with different examiners. The present study provides data to determine whether such modification is necessary. Preschool Stanford-Binets are compared with school-age WISCs. Comparing these two age levels has additional practical interest, because the WISC cannot be used before age five, so that the Stanford-Binet remains the only major preschool intelligence test.

Subjects and Procedure

Subjects (*S*s) were randomly selected from a clinic population born at Sloane Hospital for Women, previously described by Apgar et al. (**1**). At preschool age (mean = 49.4 months, sigma = 5.9), *S*s were asked to return to the hospital for Stanford-Binets. At school age (mean = 100.2 months, sigma = 6.3), *S*s were asked to reappear for WISCs. The average interval between tests was 50.8 months (sigma = 2.2).

Of 404 *S*s selected, 119 returned for both tests in response to standard mail requests. Six *S*s were excluded from the sample. Two were not testable, three had possible brain damage occurring in the intertest interval, and one had an intertest interval exceeding the mean interval by five sigmas. The final sample numbered 113, 61 males and 52 females. There were 39 white *S*s, 66 Negroes, 6 Puerto Ricans, and 2 Orientals.

One psychologist administered all Stanford-Binets, Form L, at the preschool age. Another administered all WISCs at school age. Stanford-Binet scores were withheld from the WISC examiner until testing was completed.

Results and Discussion

Intertest Correlations

The Stanford-Binet IQs correlated .67 with the WISC Full Scale IQs ($p < .01$). Though significant, the relationship is considerably lower than .85, the intertest correlation at the same age, with the same examiner. However, the .67 correlation compares favorably with previously reported correlations between preschool and school-age Stanford-Binets. The median

[1] This investigation, part of a larger research project, was supported by research grant (3B-9007) from the National Institute of Neurological Diseases and Blindness, of the National Institutes of Health, United States Public Health Service.

The authors wish to express their appreciation to William Langford for his guidance and generosity in providing the facilities to make this study possible, and to Arthur Carr, and Joseph Zubin for their helpful advice.

correlation obtained from the latter reports was found to be .74 (**3, 4, 6, 8**).

The Stanford-Binet IQs correlated .64 with the WISC Verbal Scale and .48 with the WISC Performance Scale IQs (both p's $< .01$), both lower than the correlation for the Full Scale IQs. This hierarchy of correlations agrees with previous findings.

Comparison of Mean IQs

Table 1 provides the data to compare the mean IQs for the present sample. It can be seen that the mean Stanford-Binet IQ was significantly higher than all three mean WISC IQs. This finding too agrees with previous reports.

However, in the present study, with one examiner administering all Stanford-Binets, and another all WISCs, it can be argued that the mean differences reflected examiner bias rather than intertest differences. Data were available to evaluate this alternative interpretation of the findings. If the results reflected examiner bias, one would expect lowest intertest correlations and largest intertest differences where scoring was most subjective. Since the WISC Verbal Scale entails greater subjective scoring than the Performance Scale, one would predict a lower intertest correlation and larger intertest differences for the former than the latter. Actually the reverse was found. The Verbal Scale correlated better with the Stanford-Binet and showed a smaller mean intertest difference (Table 1) than the Performance Scale. Thus, the hypothesis of examiner bias does not appear to be supported by the data.

Effect of IQ, Age, and Sex

An analysis of variance was performed to see if the observed difference between the mean Stanford-Binet and WISC Full Scale IQs was related to IQ, age, or sex. Three Stanford-Binet IQ subgroups were created, below 90, 90 to 110, and above 110. Two age subgroups were formed, below eight years and above eight. The number in each subgroup is shown in Table 2. Since the numbers were unequal, it was necessary to use the Walker-Lev (**14**) approximate method of analysis of variance.[2]

TABLE 1

COMPARISON OF MEAN STANFORD-BINET AND WISC IQS

(N = 113)

Measure	S-B	WISC		
		FS	VS	PS
Mean	104.32	98.94	100.14	97.87
SD	15.96	11.26	11.40	13.07
t:S-B vs. WISC		4.89*	3.60*	4.57*

* Significant at the .001 level

TABLE 2

EFFECT OF IQ, AGE, AND SEX ON MEAN INTERTEST DIFFERENCES

Variable	N	Mean[a]	F
S-B IQ			12.83*
<90	19	−2.68	
90–110	59	+2.39	
>110	35	+14.86	
WISC Age			1.12
>8	84	+4.76	
<8	29	+7.24	
Sex			1.50
Male	61	+3.64	
Female	52	+7.46	

[a] Minus (−) denotes WISC higher than S-B; plus (+) denotes S-B higher than WISC.

* Significant at the .001 level.

The results shown in Table 2 indicate that age and sex had no effect on intertest differences, while IQ did. Individual t tests revealed that the difference between low and average IQ levels was significant at the .05 level ($t = 2.26$), while the dif-

[2] Since some *S*s had higher Stanford-Binet IQs relative to their WISC IQs while others had higher WISC IQs, intertest differences were transformed to a unidirectional scale to calculate the analysis. The scale used assumed that zero intertest difference equals 30 points.

ferences between low and high IQ levels ($t = 6.59$) and average and high IQ levels ($t = 5.47$) were significant at the .01 level. None of the interactions was significant. The results indicate that increments in the preschool Stanford-Binet IQ increase the likelihood that it will be significantly higher than the school-age WISC IQ, and that the greatest intertest discrepancies occur at the high IQ levels.

The finding of greatest intertest differences at the high IQ levels agrees with previous research. The data on age appear to differ with the previous report (**10**) of a greater intertest discrepancy at low ages. However, since the age range of the previous study was eight years larger than that of the present study, the negative findings may be attributed to the relative homogeneity of the sample. There have been no previous reports of sex differences in relation to differences between the Stanford-Binet and the WISC.

Effect of Race and Nationality

Since the sample contained both white *S*s and Negroes, it was possible to study the effect of race on intertest differences. The results showed that both white *S*s and Negroes obtained higher Stanford-Binet IQs relative to their WISC IQs, 7.74 and 5.18 mean points, respectively. A *t* value of 1.02 indicated no significant difference between the means for both races.

Though the sample also contained six Puerto Ricans and two Orientals, their number was too small to permit comparison. However, it was necessary to demonstrate that these eight *S*s did not significantly affect the results for the remaining sample. A comparison of samples including and excluding the eight *S*s showed that both samples obtained higher Stanford-Binet IQs relative to their WISC IQs, 5.60 and 4.60 mean points, respectively. A *t* value of .70 comparing the means was not significant, indicating that the eight *S*s did not significantly affect the results.

Summary

Previous investigators have compared the Stanford-Binet and WISC IQs of *S*s retested at the same age by the same examiner. The present study attempted to ascertain whether previous findings apply to the situation where retesting occurs at different ages by different examiners.

*S*s were randomly selected from a neonatal clinic population in New York City. One psychologist administered all Stanford-Binets at preschool age; another all WISCs at school-age.

Results show that the intertest correlation is decreased from .85, for retesting at the same age by the same examiner, to .67, for retesting at different ages with different examiners. However, the .67 correlation compares favorably with similar retest findings for the Stanford-Binet itself. In other respects, results support previous findings. Highest intertest correlations were obtained for the WISC Full Scale IQ, lowest for the WISC Performance Scale IQ; the mean Stanford-Binet IQ was significantly higher than the mean WISC IQ; and greatest intertest discrepancies occurred at high IQ levels. The results appeared comparable for white *S*s and Negroes. Further, a small group of eight Puerto Rican and Oriental *S*s did not appear to affect the findings.

REFERENCES

1. Apgar, Virginia, Girdany, B. R., McIntosh, R., & Taylor, H. C. Neonatal anoxia. *Pediat.*, 1955, **15**, 653–661.
2. Arnold, F. C., & Wagner, Winifred K. A comparison of Wechsler Childrens Scale and Stanford-Binet scores for eight-and-nine year olds. *J. exp. Educ.*, 1955, **24**, 91–94.
3. Baldwin, A. L. Variation in Stanford-Binet IQ resulting from an artifact of the test. *J. Pers.*, 1948, **17**, 186–198.
4. Bradway, Katherine P. IQ constancy on the Revised Stanford-Binet. *J. genet. Psychol.*, 1944, **65**, 197–217.
5. Cohen, B. D., & Collier, Mary J. A note on the WISC and other tests of

children six to eight years old. *J. consult. Psychol.*, 1952, **16**, 226–227.

6. Diller, L., & Beechley, R. M. The constancy of the altitude. *J. clin. Psychol.*, 1951, **7**, 191–193.
7. Frandsen, Arden N., & Higginson, J. B. The Stanford-Binet and the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1951, **15**, 236–238.
8. Hilden, A. H. A longitudinal study of intellectual development. *J. Psychol.*, 1949, **28**, 187–214.
9. Holland, G. A. A comparison of the WISC and Stanford-Binet IQ's of normal children. *J. consult. Psychol.*, 1953, **17**, 147–152.
10. Krugman, Judith I., Justman, J., Wrightstone, J. W., & Krugman, M. Pupil functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children. *J. consult. Psychol.*, 1951, **15**, 475–483.
11. Kureth, Genevieve, Muhr, Jean P., & Weisberg, C. A. Some data on the validity of the Wechsler Intelligence Scale for Children. *Child Develpm.*, 1952, **23**, 281–287.
12. Mussen, P., Dean, S., & Rosenberg, Margery. Some further evidence on the validity of the WISC. *J. consult. Psychol.*, 1952, **16**, 410–411.
13. Pastovic, J. J., & Guthrie, G. M. Some evidence on the validity of the WISC. *J. consult. Psychol.*, 1951, **15**, 385–386.
14. Walker, Helen, & Lev, J. *Statistical Inference*. New York: Holt, 1953.
15. Weider, A., Noller, P. A., & Schramm, T. A. The Weschler Intelligence Scale for Children and the Revised Stanford-Binet. *J. consult. Psychol.*, 1951, **15**, 330–333.

Received July 14, 1958.

Journal of Educational Psychology
Vol. 49, No. 6, 1958

CONSERVATION OF TEACHING TIME THROUGH THE USE OF LECTURE CLASSES AND STUDENT ASSISTANTS

RUTH CHURCHILL AND PAULA JOHN

Antioch College

In 1956–7, a member of the mathematics department at Antioch College[1] became interested in contrasting what students learned when they were taught in small lecture-discussion sections with a laboratory led by the instructor and what they learned when the instructor lectured to a large class which was led in small group discussions and laboratories by a student assistant. He hoped that two kinds of savings in teaching time could be made: first, each section takes as much teaching time and time used in preparation as does a single large lecture group. Second, if upperclass students can substitute for instructors in the laboratory, the instructor is freed for additional hours by a less highly skilled person. In the experimental year, teaching by sections took 18 hours a week of faculty time; teaching by lecture and student-led laboratories took four hours a week of faculty time and ten hours of a student assistant's time.

The specific hypotheses formulated were that:

1. Students would learn skills and understandings relevant to the objectives of a course in fundamentals of mathematics, and this learning would be independent of the method of teaching the course. The specific methods contrasted were small lecture-discussion sections with a laboratory, all led by the instructor, and large lecture class, with discussions and laboratory led by an undergraduate student assistant.

2. Student attitudes towards the course would also be independent of the method of teaching, that is, equally satisfying situations could be created under both methods.

Methods

The following procedure was set up: In one division[2] the mathematics course was taught in two sections, ranging in size from 20 to 30 students, in the usual manner, three meetings a week in which lectures by the instructor were combined with questions and discussion by the students. In addition, there was a weekly laboratory session (usually an hour long), also led by the instructor. In the other division, the instructor lectured twice a week to a class of about 70 students. This group had two laboratory sessions each week, in which all discussion, questions, and help were handled by a student assistant.

Two aspects of learning in the course were selected for evaluation: background in skills and understanding of the nature of mathematics. A 126-item multiple-choice test was used to measure skills. Understanding of the nature of mathematics, considered especially important in terms of the objectives of the course, was measured by a short essay.

Student attitudes towards the course were measured by student ratings of the

[1] Acknowledgements should be made to Gustave Rabson, who initiated the experiment and taught the classes involved; to Joan Pomerantz, senior major in government and mathematics, who served as the laboratory assistant; and to Lawrence Balch, mathematics major, who served as an essay grader.

[2] At Antioch, because of the cooperative work-study plan, the student body is in two divisions, which alternate on campus, one division studying while the other is away working. Thus, the two groups, or divisions, of students in the experiment were not on campus at the same time.

instructor and by direct questions about the course. The student evaluation of the instructor employed a rating scale involving five ratings: clarity of presentation, interest in the student, arousing interest in the subject matter, making learning active, and knowledge of the subject matter. Student evaluation of the course consisted of three open-ended questions: What aspects of the course did you like most? What aspects of the course did you like least? In what ways could the course be improved?

Students took the background test and wrote essays at the beginning and the end of the course, which was 20 weeks long. The instructor was rated in the fourteenth week of the course, and the course questionnaire administered in the last week.

Unfortunately, the method used for answering and scoring the background test did not permit ascertaining its reliability easily. Students were instructed not to guess but to mark all answers possibly right. The right answers were weighted to equal the sum of the possible wrong answers; the score was the weighted sum of right answers minus one point for each wrong answer. The only available data bearing on the reliability of the test was a correlation of .74 between pre- and posttest scores for 18 students in a section of the course not in the experiment. On the whole, the test is probably sufficiently reliable to detect group differences.

Since there are no objective standards for measuring understanding of mathematics, the validity of the essay as a measure of understanding depended on the grading scale evolved and its reliability. The procedure for grading the essays was to group together all the essays, pre- and posttest, from sections and lecture class, for each of the four topics and to grade each topic separately. All identification both of student and of time was removed from the papers. Two graders were used; they had available model essays written by the instructor; and they agreed on a general definition of understanding of mathematics. The correlation between the two graders was .67. Since the sum of the two grades was used as the final grade, the reliability of the essay corrected by the Spearman-Brown formula for doubling the length is .80.

Results and Discussion

1. The results in Table 1 indicate that, in terms of pretest scores on the background test and on the essays, students taking the course in sections did not differ from those taking it in the lecture class.

2. The data in Table 1 indicate that both the sections and the lecture class gained significantly from pre- to posttest on the background test and on the essays. They further indicate that the sections did not differ from the lecture class in amount of gain.

3. When the student ratings of the instructor for the two sections are compared in Table 2 with the ratings for the lecture class, two significant differences are apparent. The large lecture class rated the instructor significantly poorer in clarity of presentation, and in general they

TABLE 1

Pre- and Posttest Scores and Gains Made by Sections and Lecture Class on the Background Test and Essays

Test	Sections A and B		Lecture class		Sign. of Diff.
	Mean	*SD*	Mean	*SD*	
Background Test	$(N = 47)$		$(N = 59)$		t
Pretest	115.6	36.8	113.9	40.3	0.6
Posttest	192.8	42.5	187.8	47.9	0.2
Gain	77.1	36.1	73.9	40.4	0.3
t_{Gain}	14.50*		13.94*		
Essay Test	$(N = 41)$		$(N = 54)$		F
Pretest	17.0	7.6	18.6	6.4	0.8
Posttest	22.7	4.6	24.9	7.1	1.6
Gain	5.7	8.0	6.3	7.7	0.1
t_{Gain}	4.5*		6.0*		

* Significant at the 1% level.

TABLE 2

STUDENT RATINGS OF THE INSTRUCTOR

Trait	Sections A and B ($N = 46$)		Lecture class ($N = 55$)		Sign. of Diff. F
	Mean	SD	Mean	SD	
Presents material clearly	2.0	1.1	3.0	1.6	12.8*
Displays interest in student	1.7	1.3	2.0	1.0	2.0
Arouses interest in subject matter	2.2	1.2	2.4	1.2	0.8
Makes learning active	2.0	1.2	2.4	1.2	2.9
Knows material	1.5	0.9	1.9	1.0	3.8
Over-all	9.5	4.2	11.8	4.6	6.9*

Note.—The lower the rating, the more favorable.

* Significant at the 1% level.

rated him slightly poorer on all traits so that the over-all rating in the lecture class is significantly poorer than that received in the sections. In both classes, however, the instructor was rated well above average when compared with a sample of the whole faculty.

Table 3 summarizes both the comments made by students on their ratings of the instructor and their answers to the threc open-ended questions used to rate the course rather than the instructor. Students in both the small sections and the lecture class responded to the questions on the most and least liked aspects of the course predominantly in terms of content. When attention is focused on the instructor rather than the course, content drops out as a relevant category. Other than this major difference, evoked by the different structuring of the two situations, the two kinds of comments were similar.

The instructor's presentation of the material was clearly the most important variable present in both kinds of comments. Presentation was commented on more often favorably and less often unfavorably in the sections than in the lecture class, significantly so in the case of unfavorable comments on the instructor. Another major factor, mentioned only unfavorably, was the rapid pace of the course. When

TABLE 3

PERCENTAGE OF STUDENTS IN SECTIONS (S) AND LECTURE (L) MAKING EACH COMMENT ON COURSE AND INSTRUCTOR

Comment	On course (N = 48S, 56L)						On instructor (N = 46S, 55L)			
	Aspects liked—				How improved?		Favorable		Unfavorable	
	Most		Least							
	S	L	S	L	S	L	S	L	S	L
Content	83	80	50	50	49	50	0	0	11	11
Presentation	50	39	17	21	17	23	78	67	15**	40**
Pace	0	0	23	23	44	45	0	0	4**	33**
Laboratories	2**	30**	2*	12*	2*	14*	0**	29**	0	2
Class size[a]	2	0	2	7	2**	20**	0	0	4	15
Examinations	8	7	4	4	0	0	4	2	2	4
Everything good			6	5	4	9				
Miscellaneous	12	8	4	2	16	20	7	0	2	5

*χ^2 yielded difference significant at 5% level.

**χ^2 yielded difference significant at 1% level.

[a] Class size too large placed in unfavorable categories.

commenting on the instructor, the lecture class made significantly more unfavorable comments on pace. The significantly greater number of unfavorable comments made by the lecture class on presentation and pace support the significantly poorer rating which they gave the instructor on clarity of presentation.

The sections and the lecture class disagreed markedly about the laboratory, which was a favorable feature for the lecture group but not mentioned by the sections. This difference can be accounted for in terms of differences in instructional procedure: for the sections the laboratory was only another meeting with the instructor while in the large class the small laboratory groups, which met with the student assistant, were a distinct feature.

In respect to the hypothesis relating to student attitudes towards the course the final position must be a qualified rejection of the null hypothesis. The lecture class was somewhat less satisfied, particularly in respect to clarity of presentation; but the lecture class commented favorably on the laboratory. The lower ratings of the instructor on clarity of presentation may have occurred because part of his function had been taken over by the laboratory assistant.

Summary

1. The problem of whether or not faculty time can be conserved through teaching in a large lecture class rather than in small sections and through replacing the instructor in the laboratory by an undergraduate student assistant was investigated by having the same instructor teach equated groups of typical students the same general education course in mathematics under two conditions: small lecture-discussion sections with a laboratory conducted by the instructor and a large lecture class with a laboratory conducted by a student assistant.

2. On pre- and posttests on a test of relevant content and on an essay graded for understanding of mathematics, the two types of classes did not differ in amount of gain and both gained significantly and substantially.

3. Although students in both types of courses rated the instructor and the course satisfactory, the lecture class was less satisfied than the sections. However, the comments in the lecture class indicated that the laboratory helped to meet student needs for discussion in which they could clarify the lecture for themselves.

Received July 17, 1958.

Journal of Educational Psychology
Vol. 49, No. 6, 1958

A STUDY OF CONSISTENT DISCREPANCIES BETWEEN INSTRUCTOR GRADES AND TERM-END EXAMINATION GRADES[1]

ELDON G. KELLY

North American Aviation, Inc., Canoga Park, California

During the past decade, a number of investigations have been made to discover the relationship of factors other than intelligence and aptitude scores to students' level of achievement. In general, students have been selected for such studies on the basis of discrepancies between their predicted level of achievement, as determined by aptitude tests and other criteria, and their actual level of achievement.

Overachieving college students have been described in one report as likely to have "less fortunate backgrounds." Factors such as social enjoyment and prestige were generally found to have influenced the underachievers' decision to attend college (**7**). Other investigators have found certain personality factors, e.g., tendencies toward maladjustment (**1**), superego status (**11**), and overconformity (**10**), to have some influence on level of achievement. Studies of the effects of remedial reading programs on achievement suggest that such programs result in improved achievement for some students (**8, 9**). Instructor grades appear to have been the only measure of achievement used in the above studies.

The purpose of the investigation to be discussed here was to discover some of the factors responsible for differences in achievement as measured by instructors' ratings and by common departmental term-end examinations. It appeared that some students in the Basic College General Education Program at Michigan State University quite consistently received a higher grade on their term-end examination than they received from their instructors in the Basic College courses, while other students seemed to be equally consistent in getting the higher of the two grades from their instructors.

The curriculum in the Basic College embodies four comprehensive areas: Communication Skills, Natural Science, Social Science, and Humanities. Each of the basics consists of three courses taken in sequence, and both instructors and students are provided a common syllabus for each course.

Students' final grades in each of the Basic College courses are derived from instructors' ratings and performance on departmental term-end examinations, each of which counts 50 per cent in the determination of the final grade.[2] Prior to conversion to the final letter grade, both instructor grades and term-end examination grades are assigned from a 15-point scale, with a score of one corresponding to F minus and a score of 15 corresponding to A plus. For each student who completes the Basic College Program, there is a record of 12 instructor grades, 12 term-end examination grades, and 12 final letter grades. The coefficient of correlation between mean instructor grades and mean term-end examination grades of all Basic College students is generally approximately .80.

[1] This study was part of a doctoral dissertation completed in 1956 under the direction of Paul L. Dressel and Walter F. Johnson, Jr.

[2] Departmental term-end examinations are multiple-choice tests constructed by the Basic College Evaluation Services, a noninstructional department which helps develop, coordinate, and administer the program of examinations and evaluation, in conjunction with the various departments involved.

Procedure

Students whose instructor grades were generally higher than their term-end examination grades would appear to be characterized by traits which enhanced their performance in the structure of classroom activities and which commended them to their instructors. Thus, the hypothesis was presented that students who generally received the higher grade from their instructors were more insecure, compulsive, conforming, and rigid than students who generally received the higher grades on the term-end examinations. The Inventory of Beliefs test was used to test this hypothesis.[3] This test consists of 120 statements with directions requesting the student to respond to each item in terms of the following key: 1. strongly agree, 2. agree, 3. disagree, and 4. strongly disagree. Since all of the statements should elicit disagreement, low scores are obtained by individuals who are characterized in terms of the above hypothesis, with the opposite being true of students obtaining high scores (**4**).

Basic College departmental term-end examinations are multiple-choice tests which are cumulative and increasingly comprehensive from term to term. These tests require a considerable amount of reading during the examination period. Thus, a second hypothesis presented was that the reading ability of students who generally received the higher grade from their instructors was inferior to that of their opposites. The Michigan State University Reading Test, designed by members of the Basic College Evaluation Services, was used to test this hypothesis. This test yields a vocabulary score, a comprehension score, and a total score.

Students whose performance on the term-end examinations was consistently short of expectations evolving from their instructor ratings might well learn to anticipate the term-end examination as a threatening, anxiety-producing experience. The Taylor Anxiety Scale was used to test an hypothesis presented with respect to this problem, with the unsurprising result that anxiety thus measured was shown to be unrelated to the problem. Beier has demonstrated, however, that induced anxiety can impair certain aspects of intellectual functioning, resulting in impaired performance on tests.[4]

The assumption underlying these hypotheses was that differences in general scholastic aptitude and intelligence were not related to the phenomenon to be studied. Nevertheless, it seemed inappropriate completely to disregard these factors, and comparisons on ACE scores were also made. The ACE, like the MSU Reading Test, is taken by all entering freshmen and it was the scores that the students in the study made on these tests at the time of enrollment which were used for the investigation.

Three groups of students were selected for the investigation of the problem: (*a*) students whose instructor grades were generally higher than their term-end examination grades (higher instructor grade group); (*b*) students whose term-end examination grades were generally the higher (higher examination grade group); and (*c*) students whose instructor grades and term-end examination grades were generally about the same (nondeviant grade group). The latter group was selected for purposes of comparison with both of the two groups above to determine if these two extreme groups were different from a nondeviant grade group as well as from

[3] Developed by the Intercollege Committee on Attitudes, Values, and Personal Adjustment: The Cooperative Study of Evaluation in General Education of the American Council on Education.

[4] The *t* technique was used in preference to the generally more appropriate analysis of variance for data of this type because of the investigator's interest in making pairwise comparisons of the groups.

each other with respect to the factors studied.

Statistical calculations for the selection of the above groups were based upon the instructor grades and term-end examination grades obtained by populations of 565 males and 469 females during their completion of the 12 Basic College courses. Other investigators have found that women tend to get higher grades than men from instructors, while men tend to get higher grades than women on standard achievement tests **(2, 12)**. To avoid this bias, means of the accumulative sums of examination and instructor grades, mean differences between the accumulative sums, and standard deviations of the differences were computed separately for men and women. Men and women thus selected were jointly assigned to their appropriate groups. (Women received both higher instructor grades and higher examination grades than men. While women's examination grades were only slightly higher than their instructor grades and only slightly higher than the men's examination grades, women's instructor grades were substantially higher than the men's.) In order to limit the study to extreme cases, only those men and women were selected whose differences between their summed examination grades and summed instructor grades placed them at least two standard deviations beyond the mean difference (E–I) between the accumulative sums of examination and instructor grades of the total male and female populations respectively.

The above method of selection identified 42 students as consistently obtaining higher grades from instructors and 54 as consistently obtaining higher grades on the term-end examination. Of these numbers, 29 students in the higher instructor grade group (14 males and 15 females) and 32 students in the higher examination grade group (20 males and 12 females) cooperated throughout the study. The nondeviant grade group was comprised of 32 students whose differences between instructor grades and examination grades placed them within one third of one standard deviation of the mean difference between the accumulative sums of examination and instructor grades.

Members of the higher instructor grade group and higher examination grade group were interviewed prior to testing. Information gained from the interviews is discussed below.

Results

After a brief, standard description of the problem, students in the higher instructor grade and higher examination grade groups were asked if they knew which category described their performance. Only one interviewee (in the higher instructor grade group) was unaware of the direction of her grades. Following the response to the above query, structuring of the interviews was restricted to the question: "How do you account for this?"

In response to the above question, 15 students in the higher instructor grade group expressed fear of the term-end examination; 11 students in the higher instructor grade group stated that too often information required on the term-end examination did not correspond to work covered in class; several labeled the examination "too ambiguous"; and some complained that both the tests as a whole and individual items were too long, requiring too much reading in the time allotted for the test.

By contrast, the comments of 25 of the 32 students in the higher examination grade group were interpreted to indicate a lack of motivation for and indifference toward the Basic College courses. Students in the higher examination grade group generally saw the disparity between their examination and instructor grades as a phenomenon of their own making, while their much more anxious opposites tended to see their circumstance as a rather threatening problem which had eluded remedy.

Mean grades presented in Table 1 below suggest, indeed, that students in the higher instructor grade group had some reason to feel threatened by the term-end examination. Their mean examination grade, about D plus, was, however, what one might expect of this group. In finding a group with higher instructor grades, one might expect to find their examination grades below average. Conversely, higher examination grades would seem to be associated with below average instructor grades. Instead, we find the average instructor grade to be the same for the two groups (C plus), and the mean examination grade of the higher examination grade group considerably above average (B plus). The extremely high coefficients of correlation between mean instructor grades and mean examination grades seen in Table 1 are artifacts of the selection method. This artifact of selection manifests itself each time both mean instructor and mean examination grades are compared with a third variable.

TABLE 1

MEAN INSTRUCTOR AND MEAN EXAMINATION GRADES, STANDARD DEVIATIONS, AND COEFFICIENTS OF CORRELATION BETWEEN MEAN INSTRUCTOR AND MEAN EXAMINATION GRADES

Deviate Groups	Mean E Grade	*SD*	Mean I Grade	*SD*	*r*
Higher I Grades	6.81	1.29	9.15	1.21	.95
Higher E Grades	11.63	1.46	9.06	1.49	.93
Non-Deviate Grades	9.36	1.46	9.44	1.46	.98

Tests of Significance of Differences

In comparing the mean ACE scores, mean reading scores, and mean Inventory of Belief scores (see Table 2), considerable differences were found to exist between the higher instructor grade group and both of the other two groups. The mean ACE scores of both the higher examination grade group and the nondeviant group were significantly higher than the mean ACE score of the higher instructor grade group. The difference between the mean ACE scores of the higher examination grade group and the higher instructor grade group was significant beyond the .001 level of confidence, while the difference between mean ACE scores of the higher instructor grade group and the nondeviant grade group was significant beyond the .01 level of confidence.[3]

Superior reading ability set the group with the higher examination grades apart from the other two groups. The mean reading score of the higher examination grade group is significantly higher than that of the nondeviant grade group beyond the .001 level of confidence, while the mean reading score of the latter group is significantly higher than that of the group with the higher instructor grades beyond the .001 level of confidence. The very small variance among the reading scores of the higher instructor grade group is one of the striking features of this group.

Mean Inventory of Belief scores revealed no difference between the higher examination grade group and the nondeviant group, but, as the data in Table 2 indicate, the mean Inventory of Belief scores of both these groups were found to be significantly higher than that of the higher instructor grade group beyond the .001 level of confidence. The group getting higher instructor grades was thus characterized as being more compulsive, insecure, rigid, and conforming.

Correlation Analysis

Estimates of the relationship of students' ACE scores, reading scores, and

TABLE 2

MEAN TEST SCORES, VARIANCES, AND TESTS OF SIGNIFICANCE OF DIFFERENCES BETWEEN EACH OF THE THREE GROUPS

Tests/All Groups	Mean	Variance	t		
			Hi'r I Gr.	Nondev. Gr.	Hi'r E Gr.
ACE					
Higher E Grades	114.06	302.06	−4.54**		
Higher I Grades	92.96	346.17		−2.97*	
Nondev. Grades	108.18	447.40			1.20
MSU Reading Test					
Higher E Grades	57.97	154.10	−7.82**		
Higher I Grades	36.93	59.37		−4.15**	
Nondev. Grades	47.81	144.34			3.30**
Inventory of Beliefs					
Higher E Grades	79.29	203.11	−3.30**		
Higher I Grades	66.17	304.49		−3.31**	
Nondev. Grades	78.34	116.91			.30

* Significant beyond the .01 level of confidence.

** Significant beyond the .001 level of confidence.

TABLE 3

CORRELATION COEFFICIENTS SHOWING RELATIONSHIP BETWEEN STUDENTS' TEST SCORES AND MEAN EXAMINATION AND INSTRUCTOR GRADES

Variables Correlated	r_{bis}			r_{tri}[a]
	Higher E Grades	Higher I Grades	Non-deviate Grades	
ACE & X E Grades	.35	.77	.25	
ACE & X I Grades	.37	.86	.27	
ACE Scores for All Grps and (E − I)				.45
Reading & X E Grades	.55	.57	.45	
Reading & X I Grades	.55	.61	.45	
Reading Scores for All Grps and (E − I)				.68
IB & X E Grades	.15	−.10	.00	
IB & X I Grades	.22	−.15	.05	
IB Scores for All Groups and (E − I)				.52

[a] Triserial correlation coefficients showing relationships of test scores of all groups to differences between examination and instructor grades (E − I).

Inventory of Belief scores to their mean instructor and mean examination grades (see Table 3) resulted in coefficients of correlation which, in general, require little comment.

The magnitudes of the coefficients obtained in estimating the relationship of ACE scores to mean examination and mean instructor grades for the higher instructor grade group, .77 and .86 respectively, are considerably greater than those obtained for the other two groups and greater, too, than customarily found. Within this group, apparently, both the examination and the instructor ranked the students quite consistently in relation to ability. There are personal qualities at work, however, which seem to commend the student to the instructor, resulting in higher grades from instructors.

The values obtained in estimating the relationship of reading scores and Inventory of Belief scores to mean instructor and mean examination grades are similar to those values usually found in using these instruments. The Inventory of Beliefs typically yields a rather wide range of scores.

Analysis of the relationship of the *differences* between students' examination and instructor grades to test scores revealed rather substantial evidence that for these groups higher aptitude scores, reading scores, and Inventory of Belief scores were positively related to tendencies to get the higher grade on the term-end examination. Jaspen's formula for triserial correlation was used to determine the relationship of difference between examination and instructor grades to test scores (**5**). A triserial coefficient of .45 was obtained in estimating the relationship of difference between examination and instructor grades to ACE scores; a coefficient of .68 was obtained in correlating reading scores with these grade differences; and a coefficient of .52 was obtained in estimating the relationship of Inventory of Belief scores to differences between examination and instructor grades.

Discussion

The mean instructor and examination grades of the higher instructor grade group were in keeping with the investigator's expectations, i.e., average grades from instructors and below average grades on the term-end examination. Expectations of the converse for the higher examination grade group were not supported by the results of the study. This group's very high mean ACE score and reading score also suggest the unlikelihood of finding many of the expected variety in the group.

Considered from the point of view of ability to achieve, the evidence suggests that the higher instructor grade group received higher grades from their instructors than they should have, while the higher examination grade group received lower instructor grades than they should have. The superiority of the nondeviant grade group to the higher instructor grade group in general aptitude and reading ability also raises a question about the similarity of the instructor grades of these two groups. Again, the evidence seems to force the conclusion that students who were characterized as being more conforming, compulsive, rigid, and insecure received higher grades from their instructors than would be expected of them on the basis of ability alone. The information obtained in interviews suggests that the average instructor grades obtained by the higher examination group must be explained in terms of a lack of motivation for and indifference toward the Basic College courses.

No thought was given to determining which of these two groups, higher instructor grade group or higher examination grade group, were the overachievers and which the underachievers. With respect to instructor grades, the higher examination grade group could be called underachievers; but, in general, they did make high term-end examination grades, thus demonstrating a high degree of mastery of their subjects. Conversely, the higher instructor grade group could be called overachievers because their instructor grades appeared to be higher than warranted by their ability; but this group's examination grades do not indicate mastery of the subject, or overachievement. The results of this study suggest that what has often been called over or underachievement may in some cases have been a function of the *method* of measuring achievement. In such cases a student's grades might nðt be an accurate description of his relative mastery of the subject.

Summary

The purpose of this investigation was to discover some of the factors which differentiate students whose instructor grades were consistently higher than their grades on departmental term-end examinations from students who consistently got the higher grades on their departmental term-end examinations. Students who consistently received the higher grade from instructors were found to receive average instructor grades and below average grades on the term-end examinations, while students who consistently received the higher grade on the term-end examinations were found to have superior examination grades and average instructor grades. Aptitude scores and reading scores of students who received the higher grades from instructors were found to be significantly lower, beyond the .001 level of confidence, than those of their opposites in achievement. The Inventory of Beliefs test characterized the higher instructor grade group as being more compulsive, conforming, rigid, and generally insecure than their opposites.

REFERENCES

1. Altus, W. D. A college achiever and non-achiever scale from the Minnesota Multiphasic Personality Inventory. *J. appl. Psychol.*, 1948, **38**, 385–395.
2. Anastasi, Anne, & Foley, J. P.; *Di ferential psychology*, New York: M millan, 1949.
3. Beier, E. G. The effect of indi anxiety on some aspects of intel tual functioning: A study of the lationship between anxiety and r gidity; *Amer. Psychologist*, 1949, 4 273–274.
4. Dressel, P. L., & Mayhew, L. B. *General education: Explorations in evaluation*, Washington, D. C.: American Council on Education, 1954.
5. Jaspen, N. Serial correlation, *Psychometrika*, 1946, **11**(1), 23–30.
6. McQuarry, J. P. Some differences between over-achievers and under-achievers. *Educ. Adm. Superv.*, 1954, **40**, 117–120.
7. McQuarry, J. P., & Trux, W. E. An underachievement scale, *J. educ. Res.*, 1955, **48**, 393–399.
8. Mouly, G. J. A study of the effects of a remedial reading program on academic grades at the college level. *J. educ. Psychol.* 1952, **45**, 459–566.
9. Preston, R. G., & Botel, M. Relatio of reading skill and other factors t the achievement of 2048 college students. *J. exp. Educ.*, 1952, **20**, 363–371.
10. Rust, R. M. & Ryan, F. J. Relationships of some Rorschach variables to academic behavior. *J. Pers.*, 1953, **21**, 441–456.
11. Ryan, F. J. Personality differences between under- and over-achievers in college. Unpublished doctoral dissertation, Columbia Univer., 1951. (Dissertation abstract)
12. Stephens, J. M. *Educational psychology*, New York: Holt, 1951.

Received August 4, 1958.